

# Multi-Root I/O Virtualization and Sharing

## Revision 0.7

June 8, 2007

---

REVISION	REVISION HISTORY	DATE
0.5	Initial release	11/13/2006
0.7	Methods & Functions Complete	6/8/2007

PCI-SIG disclaims all warranties and liability for the use of this document and the information contained herein and assumes no responsibility for any errors that may appear in this document, nor does PCI-SIG make a commitment to update the information contained herein.

Contact the PCI-SIG office to obtain the latest revision of the specification.

Questions regarding this document or membership in PCI-SIG may be forwarded to:

**Membership Services**

<http://www.pcisig.com>

E-mail: [administration@pcisig.com](mailto:administration@pcisig.com)

Phone: 503-291-2569

Fax: 503-297-1090

**Technical Support**

Technical support for this specification is available to members. For information, please visit: [http://www.pcisig.com/developers/technical\\_support](http://www.pcisig.com/developers/technical_support).

**DISCLAIMER**

This document is provided “as is” with no warranties whatsoever, including any warranty of merchantability, non-infringement, fitness for any particular purpose, or any warranty otherwise arising out of any proposal, specification, or sample. PCI-SIG disclaims all liability for infringement of proprietary rights, relating to use of information in this specification. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

PCI Express, PCIe, PCI-X, and PCI-SIG are trademarks of PCI-SIG.

All other product names are trademarks, registered trademarks, or service marks of their respective owners.

Copyright © 2007 PCI-SIG  
All rights reserved.

# Contents

<b>1. ARCHITECTURAL OVERVIEW .....</b>	<b>17</b>
1.1. HOW DOES MR-IOV WORK? .....	19
1.1.1. MRA Components .....	24
1.1.1.1. Multi-Root Aware Root Port (MRA RP) .....	24
1.1.1.2. Multi-Root Aware PCIe Device (MRA PCIe Device) .....	25
1.1.1.3. Multi-Root PCI Manager (MR-PCIM) .....	26
1.1.1.4. Multi-Root Aware PCIe Switch (MRA PCIe Switch) .....	27
1.1.2. MR Initialization Overview .....	29
1.1.3. MR Transaction Encapsulation Overview .....	29
1.1.4. MR Congestion Management Overview .....	29
1.1.5. MR Error and Event Handling Overview .....	29
1.1.6. MR-IOV and ARI (Alternative Routing Identifier) .....	29
1.1.7. MR-IOV Relationship to SR-IOV and ATS .....	29
1.2. OVERVIEW OF MR TRANSACTION LAYER .....	30
<b>2. MR PROTOCOL CHANGES .....</b>	<b>34</b>
2.1. MR LINK PROTOCOL NEGOTIATION .....	34
2.1.1. MR Link Protocol Negotiation .....	38
2.1.2. MR Flow Control Initialization Protocol .....	39
2.1.2.1. MR Flow Control DLLP Encoding .....	40
2.1.2.2. MR Flow Control Initialization State Machine Rules .....	45
2.2. TLP PREFIX TAGGING .....	49
2.2.1. MR Switch Transaction Layer Processing .....	50
2.2.2. MR Device Transaction Layer Processing .....	52
2.2.2.1. Receiving TLPs .....	52
2.2.2.2. Transmitting TLPs .....	53
2.2.3. Global Key Processing .....	54
2.2.4. MR TLP Dataflow Examples .....	55
2.3. PER-VH RESET .....	56
2.3.1. Per-VH Reset Example .....	57
2.3.2. RESET DLLP Format .....	60
2.3.3. RESET DLLP Processing .....	60
2.3.3.1. Upstream State Machine .....	61
2.3.3.2. Downstream State Machine .....	62
2.3.3.3. Reset DLLP Reliability .....	64
2.3.3.4. Flow Control and Reset / DL_DOWN .....	65
2.4. MR FLOW CONTROL .....	65
2.4.1. FC Information Tracked by Transmitter .....	65
2.4.2. Information Tracked by Receiver .....	67
2.5. MR MESSAGE PROCESSING .....	67
2.5.1. Interrupts .....	67
2.5.1.1. INTx Device Processing .....	67
2.5.1.2. INTx Switch Processing .....	68

2.5.1.3.	INTx Root Port Processing.....	68
2.5.2.	PME Turn Off Processing.....	68
2.5.3.	PM_PME Processing.....	68
2.6.	MISCELLANEOUS CHANGES .....	68
2.7.	MISCELLANEOUS NON-CHANGES .....	69
<b>3.</b>	<b>INITIALIZATION AND RESOURCE ALLOCATION .....</b>	<b>70</b>
3.1.	MR TOPOLOGY INITIALIZATION .....	70
3.1.1.	Initial State after Fundamental Reset .....	71
3.1.1.1.	PCIM Capable Switch Ports.....	71
3.1.1.2.	Non-PCIM Capable Switch Ports.....	72
3.1.1.3.	Non-PCIe Switch Management Ports.....	72
3.1.1.4.	Initial State Example .....	74
3.1.2.	Initial MR-PCIM Location Policy .....	75
3.1.2.1.	Initial MR-PCIM Location Example.....	75
3.1.3.	Topology Discovery .....	75
3.1.3.1.	MR-PCIM Topology Discovery Example .....	76
3.1.4.	Component Discovery.....	78
3.1.4.1.	Component Discovery Example.....	79
3.1.5.	VH and VF Mapping Policy.....	80
3.1.5.1.	Example VH and VF Mapping Policy .....	80
3.1.6.	VH and VF Mapping Implementation.....	81
3.1.6.1.	Example Topology: Switch Implementation.....	81
3.1.6.2.	Example Topology: Device Implementation .....	85
3.1.7.	MR-PCIM Failover.....	87
3.2.	MR DEVICE INITIALIZATION.....	88
3.2.1.	Enabling MR Operation.....	89
3.2.2.	Managing Flow Control .....	90
3.2.3.	Managing VF Mapping.....	91
3.2.4.	Managing VF Migration.....	93
3.2.4.1.	VF Migration Initial State .....	95
3.2.4.2.	VF Migration Reinitialization .....	96
<b>3.3.</b>	<b>MR ROOT PORT INITIALIZATION .....</b>	<b>97</b>
<b>3.3.1.</b>	<b>Example MR Root Port Topology.....</b>	<b>98</b>
<b>4.</b>	<b>CONFIGURATION.....</b>	<b>99</b>
4.1.	CONFIGURATION FIELD SUMMARY .....	100
4.2.	DEVICE CONFIGURATION SPACE.....	106
4.2.1.	Device MR-IOV Extended Capability.....	108
4.2.1.1.	MR-IOV Extended Capability Header (00h) .....	110
4.2.1.2.	MR-IOV Capabilities (04h).....	111
4.2.1.3.	MR-IOV Control (08h) .....	112
4.2.1.4.	MR-IOV Status (0Ch) .....	114
4.2.1.5.	MR-IOV VH Counts (10h).....	115
4.2.1.6.	Function Table Offset (14h) .....	115
4.2.1.7.	MVF and LVF Sizes (18h).....	116
4.2.1.8.	LVF Table Offset (1Ch).....	117

4.2.1.9.	VL Arbitration Capability and Status (20h) .....	118
4.2.1.10.	VL Arbitration Control (24h) .....	120
4.2.1.11.	VL Arbitration Table Offset (28h) .....	121
4.2.1.12.	MR Error Status (2Ch) .....	122
4.2.1.13.	MR Error Control (2Eh) .....	122
4.2.1.14.	MR Error Log (30h to 40h) .....	122
4.2.1.15.	Statistics Capability and Control (44h to 50h) .....	122
4.2.2.	<i>Device VL Arbitration Table</i> .....	122
4.2.3.	<i>LVF Table</i> .....	123
4.2.3.1.	LVF Table Entry .....	123
4.2.4.	<i>Function Table</i> .....	124
4.2.4.1.	Function Capability (00h and 04h) .....	126
4.2.4.2.	Function Control (08h to 10h) .....	127
4.2.4.3.	Function Status (14h) .....	131
4.2.4.4.	Function VC to VL Map (18h and 1Ch) .....	132
4.2.4.5.	Function VC Resource (20h to 2Ch) .....	134
4.2.4.6.	Function Table MFVC Resource Status (30h to 3Ch) .....	134
4.2.4.7.	Function Interrupt Bitmap (minus 20h) .....	135
4.2.5.	<i>Misc. Device Configuration Space Requirements</i> .....	135
4.2.5.1.	BIST (Device) .....	135
4.3.	SWITCH CONFIGURATION SPACE .....	136
4.3.1.	<i>Switch MR-IOV Extended Capability</i> .....	138
4.3.1.1.	Switch MR-IOV Extended Capability Header (00h) .....	140
4.3.1.2.	Switch MR-IOV Capability (04h) .....	140
4.3.1.3.	Switch MR-IOV Control (08h) .....	141
4.3.1.4.	Switch MR-IOV Status (0Ch) .....	142
4.3.1.5.	MR-IOV This Bridge Map (10h) .....	143
4.3.1.6.	Watchdog Timer Control (14h) .....	143
4.3.1.7.	Authorization (18h) .....	144
4.3.1.8.	Port Table Entry Size / Num Port Entries (1Ch) .....	145
4.3.1.9.	Port Table Offset (20h) .....	145
4.3.1.10.	VS Table Entry Size / Num VS Table Entries (24h) .....	146
4.3.1.11.	VS Table Offset (28h) .....	146
4.3.1.12.	VS Bridge Table Entry Size / Num VS Bridge Table Entries per VS (2Ch) 147	
4.3.1.13.	VS Bridge Table Offset (30h) .....	147
4.3.1.14.	Statistics Capability and Control (30h to 3Ch) .....	148
4.3.2.	<i>Switch VS Authorization Bitmap</i> .....	148
4.3.3.	<i>Switch Port Table</i> .....	148
4.3.3.1.	Port Capability (00h) .....	150
4.3.3.2.	Port Control (04h and 08h) .....	151
4.3.3.3.	Port Status (0Ch) .....	154
4.3.3.4.	Link Partner Training Status (10h) .....	155
4.3.3.5.	VL Arbitration Capability and Status (14h) .....	157
4.3.3.6.	VL Arbitration Control (18h) .....	159
4.3.3.7.	VL Arbitration Table Offset (1Ch) .....	160

4.3.3.8.	MR Error Status (20h).....	160
4.3.3.9.	MR Error Control (22h) .....	160
4.3.3.10.	MR Error Log (24h to 34h) .....	160
4.3.3.11.	PCI Bridge Control (N+00h).....	161
4.3.3.12.	PCIe Capability Structure (N+02h).....	161
4.3.3.13.	Port Interrupt Status Bitmap (minus 20h) .....	164
4.3.4.	<i>Switch VL Arbitration Table .....</i>	<i>164</i>
4.3.5.	<i>Switch VS Table .....</i>	<i>165</i>
4.3.5.1.	VS Capability and Status (00h) .....	165
4.3.5.2.	VS Control (04h) .....	166
4.3.5.3.	VS Bridge Interrupt Status (08h to ??h).....	167
4.3.5.4.	VS Interrupt Status Bitmap (minus 20h) .....	168
4.3.6.	<i>Switch VS Bridge Table .....</i>	<i>168</i>
4.3.6.1.	VS Bridge Capability and Status (00h) .....	171
4.3.6.2.	VS Bridge Control (04h and 08h) .....	173
4.3.6.3.	VC ID to VL Map (0Ch and 10h) .....	176
4.3.6.4.	VC Resource Fields (14h to 20h).....	178
4.3.6.5.	Hot-Plug Virtual Signals Interface (24h and 28h).....	179
4.3.7.	<i>Misc. Switch Configuration Space Requirements.....</i>	<i>182</i>
4.3.7.1.	ARI Support .....	182
4.3.7.2.	BIST (switch) .....	182
4.4.	VL ARBITRATION TABLE.....	183
4.5.	PERFORMANCE MONITORING AND STATISTICS COLLECTION .....	184
4.5.1.	<i>Configuration Space Fields .....</i>	<i>185</i>
4.5.1.1.	Statistics Capability (+00h) .....	186
4.5.1.2.	Statistics Block Start / Busy (+04h) .....	187
4.5.1.3.	Statistics Descriptor Table Offset (+08h).....	188
4.5.1.4.	Statistics Block Table Offset (+0Ch) .....	189
4.5.2.	<i>Statistics Descriptor Table.....</i>	<i>189</i>
4.5.2.1.	Standard Statistics .....	191
4.5.2.2.	Standard Filters.....	193
4.5.3.	<i>Statistics Block Table.....</i>	<i>195</i>
4.5.3.1.	Statistics Block Capability (00h).....	195
4.5.3.2.	Statistics Table Offset (04h) .....	196
4.5.3.3.	Statistics Wait Time (08h).....	196
4.5.3.4.	Statistics Count Time (0Ch) .....	197
4.5.4.	<i>Statistics Counter Table.....</i>	<i>197</i>
4.5.4.1.	Statistics Capability and Control (00h) .....	198
4.5.4.2.	Statistics Filter Enable and Control (04h) .....	199
4.5.4.3.	Statistics Counter Low (08h).....	199
4.5.4.4.	Statistics Counter High (0Ch) .....	199
<b>5.</b>	<b>ERROR HANDLING .....</b>	<b>200</b>
5.1.	PCIe ERROR MAPPING TO MR .....	200
5.2.	MR ERRORS.....	203
<b>6.</b>	<b>HOT PLUG.....</b>	<b>205</b>

6.1.	MRA SWITCH .....	205
6.1.1.	PCI Express Capability: Slot Capability Register.....	206
6.1.2.	PCI Express Capability: Slot Control Register .....	208
6.1.3.	PCI Express Capability: Slot Status Register.....	210
6.1.4.	PCI Express Capability: Device Capabilities Register .....	212
6.1.5.	Virtual Hot-Plug Signals Interface Registers .....	212
6.1.6.	Physical Slot Registers.....	213
6.1.7.	Physical Hot-Plug Signals Interface.....	214
6.2.	VIRTUAL DEVICE MIGRATION .....	214
6.3.	BASE PCI EXPRESS DEVICE MIGRATION .....	214
<b>7.</b>	<b>POWER MANAGEMENT .....</b>	<b>215</b>
7.1.	OVERVIEW .....	215
7.2.	VIRTUAL D-STATE.....	215
7.3.	LINK POWER STATES .....	215
7.4.	MULTI-ROOT ASPM.....	216
7.5.	SLOT CLOCK AND COMMON CLOCK CONFIGURATION .....	216
7.6.	MULTI-ROOT WAKE-UP .....	216
7.6.1.	PME Triggers Beacon / Wake#.....	217
7.6.2.	Beacon / Wake# Triggers MSI.....	217
7.6.3.	Beacon / WAKE# Triggers Beacon / WAKE#.....	218
7.7.	MULTI-ROOT PME TURN OFF .....	218
7.8.	MULTI-ROOT POWER CONTROLLER.....	218
7.9.	MULTI-ROOT POWER BUDGETING .....	219
<b>8.</b>	<b>CONGESTION MANAGEMENT .....</b>	<b>220</b>
8.1.	OVERVIEW .....	220
8.2.	CONGESTION ISOLATION.....	221
8.2.1.	Virtual Links.....	221
8.2.1.1.	Virtual Link and Virtual Hierarchy Identification .....	222
8.2.1.2.	VL and VC Configuration .....	223
8.2.1.3.	VC to VL Mapping.....	224
8.2.1.4.	Arbitration .....	225
8.2.2.	Bypass Queues .....	227
8.2.3.	Flow Control Rules .....	228
8.3.	PERFORMANCE MONITORING AND STATISTICS COLLECTION .....	230

## Figures

Figure 1-1:	Generic Server Blade Configuration.....	17
Figure 1-2:	Example Multi-Root Topology .....	30
Figure 1-3:	Example Multi-Root Topology as viewed from Host A .....	32
Figure 1-4:	Example Multi-Root Topology as viewed from Host C .....	33
Figure 2-1:	MRInit DLLP Format .....	34
Figure 2-2:	MR to MR Initialization Sequence.....	36

Figure 2-3:	MR to Base Initialization .....	36
Figure 2-4:	MR Data Link Control and Management State Machine (MR-DLCMSM).....	37
Figure 2-5:	MR InitFC State Machine.....	40
Figure 2-6:	MRUpdateFC Header DLLP.....	41
Figure 2-7:	MRInitFC1_VL Header DLLP.....	42
Figure 2-8:	MRInitFC1_VH Header DLLP .....	42
Figure 2-9:	MRInitFC2_VL Header DLLP .....	42
Figure 2-10:	MRInitFC2_VH Header DLLP .....	43
Figure 2-11:	MRUpdateFC Data DLLP.....	43
Figure 2-12:	MRInitFC1_VL Data DLLP .....	43
Figure 2-13:	MRInitFC1_VH Data DLLP.....	44
Figure 2-14:	MRInitFC1_VL Data DLLP .....	44
Figure 2-15:	MRInitFC2_VH Data DLLP.....	44
Figure 2-16:	TLP Prefix Header Location .....	49
Figure 2-17:	TLP Prefix Header Layout.....	50
Figure 2-18:	MR Dataflow Examples .....	55
Figure 2-19:	Reset DLLP Example: Topology.....	57
Figure 2-20:	Reset DLLP Example: Host A's View .....	58
Figure 2-21:	Reset DLLP .....	60
Figure 2-22:	Upstream Link Partner RESET SM .....	62
Figure 2-23:	Downstream Link Partner RESET SM.....	64
Figure 3-1:	Example MR Topology .....	74
Figure 3-2:	Example MR Topology with Initial Link Directions .....	76
Figure 3-3:	Example MR Topology with Component Discovery Details.....	79
Figure 3-4:	Example MR Topology: RP 0 View .....	81
Figure 3-5:	Example MR Topology: RP 1 View .....	81
Figure 3-6:	Example MR Topology: RP 2 View .....	82
Figure 3-7:	Example MR Topology: RP 3 View .....	82
Figure 3-8:	Example MR Topology: Device L PF / VF Mapping.....	86
Figure 3-9:	Example MR Topology: Device M VF Mapping .....	87
Figure 3-10:	Example Mapping of VFs .....	92
Figure 3-11:	VF Migration State Diagram.....	94
Figure 3-12:	Initial VF State .....	96
Figure 3-13:	Example Topology with MR Root Complex .....	98
Figure 4-1:	MR Device Configuration Space.....	108
Figure 4-2:	Device MR-IOV Capability .....	110
Figure 4-3:	LVF Table.....	123
Figure 4-4:	Device Function Table .....	126
Figure 4-5:	Switch Mapping Tables.....	137
Figure 4-6:	Switch MR-IOV Capability Diagram.....	140
Figure 4-7:	Switch Port Table .....	149
Figure 4-8:	Port Interrupt Status Bitmap .....	164
Figure 4-9:	VS Table.....	165
Figure 4-10:	VS Interrupt Status Bitmap .....	168
Figure 4-11:	VS Bridge Table .....	169
Figure 4-12:	Example VL Arbitration Table with 32 Phases .....	183
Figure 4-13:	Performance Monitoring and Statistics Collection Tables .....	185
Figure 6-1:	Slot Capabilities Register (PCIe Figure 7-18) .....	206



Figure 6-2:	Slot Control Register (PCIe Figure 7-19).....	208
Figure 6-3:	Slot Status Register (PCIe Figure 7-20).....	210
Figure 6-4:	PCI Express Capabilities Register (PCIe Figure 7-11).....	212
Figure 7-1:	Multi-Root Wake-Up Scenarios.....	217
Figure 8-1:	(VH, VC) to VL Mapping .....	222
Figure 8-2:	MRA Arbitration Model.....	226
Figure 8-3:	Logical Queuing Structure Associated with a VL Receiver.....	228
Figure 8-4:	Statistics Collection Process.....	230

## Tables

Table 2-1:	MRInit DLLP Fields.....	35
Table 2-2:	MR Flow Control DLLP Fields.....	45
Table 2-3:	Reset DLLP Example: Events and Actions.....	59
Table 3-1:	Port Types – Example MR Topology .....	74
Table 3-2:	Example MR Topology VH and VF Mapping Policy .....	80
Table 3-3:	Example Topology: Switch A VS Bridge Table Contents .....	83
Table 3-4:	Example Topology: Switch B VS Bridge Table Contents.....	84
Table 3-5:	Valid MR State Transitions for VF Migration .....	94
Table 3-6:	Example MR Root Topology: RP Associations .....	98
Table 4-1:	MR-IOV Fields.....	100
Table 4-2:	Device MR-IOV Extended Capability Header.....	110
Table 4-3:	MR-IOV Capabilities.....	111
Table 4-4:	Device MR-IOV Control.....	112
Table 4-5:	Device MR-IOV Status .....	114
Table 4-6:	Device MR-IOV VH Counts .....	115
Table 4-7:	Device Function Table Offset .....	115
Table 4-8:	VF MVF Region.....	116
Table 4-9:	LVF Table Offset.....	117
Table 4-10:	Device VL Arbitration Capability and Status .....	118
Table 4-11:	Device VL Arbitration Control.....	120
Table 4-12:	Device VL Arbitration Table Offset .....	121
Table 4-13:	Device MR Error Status .....	122
Table 4-14:	Device MR Error Control.....	122
Table 4-15:	LVF Table Entry.....	123
Table 4-16:	Function Capability 1 (00h).....	126
Table 4-17:	Function Capability 2 (04h).....	126
Table 4-18:	Function Control 1 (08h) .....	127
Table 4-19:	Function Control 2 (0Ch).....	129
Table 4-20:	Function Control 3 (10h) .....	130
Table 4-21:	Function Status .....	131
Table 4-22:	Function Table VC to VL Map 1 (VC Capability) .....	132
Table 4-23:	Function Table VC to VL Map 2 (VC Capability) .....	133
Table 4-24:	Function Table VC Resource State.....	134
Table 4-25:	VH Table MFVC Resource State.....	135

Table 4-26:	Switch MR-IOV Extended Capability Header.....	140
Table 4-27:	Switch MR-IOV Capability Bits.....	140
Table 4-28:	Switch MR-IOV Control Bits.....	141
Table 4-29:	Switch MR-IOV Status Bits.....	142
Table 4-30:	Switch MR-IOV This Bridge Map.....	143
Table 4-31:	Switch MR-IOV Authorization Control.....	144
Table 4-32:	Switch MR-IOV Authorization Control.....	144
Table 4-33:	Switch Port Table Sizes.....	145
Table 4-34:	Switch Port Table Offset.....	145
Table 4-35:	Switch VS Table Sizes.....	146
Table 4-36:	Switch VS Table Offset.....	146
Table 4-37:	Switch VS Bridge Table Sizes.....	147
Table 4-38:	Switch VS Bridge Table Offset.....	147
Table 4-39:	Switch Port Capability.....	150
Table 4-40:	Switch Port Control1.....	151
Table 4-41:	Switch Port Control2.....	152
Table 4-42:	Switch Port Status.....	154
Table 4-43:	Switch Link Partner Training Status.....	155
Table 4-44:	Switch Link Partner Training Status – MRInit DLLP Bits.....	156
Table 4-45:	Switch VL Arbitration Capability and Status.....	157
Table 4-46:	Switch VL Arbitration Control.....	159
Table 4-47:	Switch VL Arbitration Table Offset.....	160
Table 4-48:	Switch MR Error Status.....	160
Table 4-49:	Switch MR Error Control.....	160
Table 4-50:	PCI Bridge Control.....	161
Table 4-51:	Port PCIe Capability Structure.....	162
Table 4-52:	Switch VS Capability and Status.....	165
Table 4-53:	Switch VS Capability and Status.....	166
Table 4-54:	Switch VS Bridge Capability and Status.....	171
Table 4-55:	Switch VS Bridge Control 1.....	173
Table 4-56:	Switch VS Bridge Control 2.....	175
Table 4-57:	Switch VS Bridge VC ID to VL Map 1.....	176
Table 4-58:	Switch VS Bridge VC ID to VL Map 2.....	177
Table 4-59:	VC Resource State.....	179
Table 4-60:	Virtual Hot-Plug Signals Interface 1.....	180
Table 4-61:	Hot-Plug Signals Interface 2.....	181
Table 4-62:	Definition of the 4-bit Entries in the VL Arbitration Table.....	183
Table 4-63:	Length of the VL Arbitration Table.....	184
Table 4-64:	Statistics Table Sizes.....	186
Table 4-65:	Statistics Start / Busy.....	187
Table 4-66:	Statistics Descriptor Table Offset.....	188
Table 4-67:	Statistics Block Table Offset.....	189
Table 4-68:	Statistics Descriptor Table Entry.....	190
Table 4-69:	Standard Statistics.....	191
Table 4-70:	TLP Filters.....	193
Table 4-71:	Credit Filters.....	194
Table 4-72:	DLLP Filters.....	195

Table 4-73:	Statistics Block Capability.....	196
Table 4-74:	Statistics Table Offset.....	196
Table 4-75:	Statistics Wait Time .....	196
Table 4-76:	Statistics Count Time .....	197
Table 4-77:	Statistics Capability and Control.....	198
Table 4-78:	Statistics Filter Enable and Control .....	199
Table 4-79:	Statistics Counter Low .....	199
Table 4-80:	Statistics Counter High .....	199
Table 5-1:	Physical Layer Error List.....	200
Table 5-2:	Data Link Layer Error List .....	200
Table 5-3:	Transaction Layer Error List.....	201
Table 5-4:	MR Error List .....	203
Table 6-1:	Virtual Mapping: PCIe Slot Capabilities Register.....	206
Table 6-2:	Virtual Mapping: PCIe Slot Control Register .....	208
Table 6-3:	Virtual Mapping: PCIe Slot Control Register .....	210
Table 6-4:	Virtual Mapping: PCIe Capabilities Register.....	212



## Objective of the Specification

The purpose of this document is to specify PCI I/O virtualization and sharing technology. The specification is focused on multi-root topologies, e.g. a server blade enclosure that uses a PCI Express Switch-based topology to connect server blades to PCI Express Devices or PCI Express-to-PCI Bridges and enable the leaf Devices to be serially or simultaneously shared by one or more server blades.

This document is to be used in conjunction with, and does not supersede, the terms and conditions specified in the PCI-SIG™ Trademark and Logo Usage Guidelines document.

## Document Organization

Chapter 1 specifies.

Chapter 2 specifies.

## Documentation Conventions

### Capitalization

Some terms are capitalized to distinguish their definition in the context of this document from their common English meaning. Words not capitalized have their common English meaning. When terms such as “memory write” or “memory read” appear completely in lower case, they include all transactions of that type.

Register names and the names of fields and bits in registers and headers are presented with the first letter capitalized and the remainder in lower case.

### Numbers and Number Bases

Hexadecimal numbers are written with a lower case “h” suffix, e.g., 0FFFFh and 80h. Hexadecimal numbers larger than four digits are represented with a space dividing each group of four digits, as in 1E FFFF FFFFh. Binary numbers are written with a lower case “b” suffix, e.g., 1001b and 10b. Binary numbers larger than four digits are written with a space dividing each group of four digits, as in 1000 0101 0010b.

All other numbers are decimal.

### Reference Information

Reference information is provided in various places to assist the reader and does not represent a requirement of this document. Such references are indicated by the abbreviation “(ref).” For example, in some places, a clock that is specified to have a minimum period of 400 ps also includes the reference information maximum clock frequency of “2.5 GHz (ref).” Requirements of other specifications also appear in various places throughout this document and are marked as reference information. Every effort has been made to guarantee that this information accurately reflects the referenced document; however, in case of a discrepancy, the original document takes precedence.

## Implementation Notes

Implementation Notes should not be considered to be part of this specification. They are included for clarification and illustration only. Implementation Notes within this document are enclosed in a box and set apart from other text.

## Terms and Abbreviations

Base Function (BF)	Function used to manage the MR features of an MR Device.
IO Virtualization (IOV)	The capability for a single PCIe component to be used by more than one SI.
Switch Management Port	A connection to a MR Switch that can be used to manage a MR topology. Switch Management Ports may be actual PCIe Ports or may be Vendor Specific non-PcIe Ports.
MR - Multi-Root Topology	A PCIe topology that interconnects more than one Root Port through multi-root aware Switches.
MR PCIM	Multi-Root PCIM
MRA	Multi-Root Aware – A PCIe component that supports the multi-root extensions defined in this specification.
MRA VF	Virtual Function (VF) in an MRA Device
MR Enabled Link	PCIe link using the Multi-Root Encapsulated Link Protocol.
MR Egress	Point in a PCIe fabric that a TLP exits the MR Fabric
MR Fabric	Subset of a PCIe fabric containing MR Enabled Links and connected MRA components.
MR Ingress	Point in a PCIe fabric that a TLP enters the MR Fabric.
PCI Bus Memory Address	Refers to the address portion of a PCI Memory Transaction
PCI Manager (PCIM)	Software that enumerates and configures an IOV topology.
PCIe	PCI Express
Physical Address	The address used by the system memory controller to access system memory.
Physical Function (PF)	An IOV-capable Function per the <i>Single Root IOV Specification</i> . In MR, PFs exist within Virtual Hierarchies (VHs).
RC	Root Complex per the <i>PCI Express Base Specification</i>
RP	Root Port per the <i>PCI Express Base Specification</i>
System Image (SI)	A software component running on a Virtual System to which specific virtual and physical Devices can be assigned. Specification of the behavior and architecture of an SI is outside the scope of this specification. Examples of SIs include guest operating systems and

shared/non-shared protected domain Device drivers.

SR PCIM	Single Root PCIM
Virtual Device	Collection of PFs and VFs that operate as a Device within a VH.
Virtual Function (VF)	An IOV-capable Function per the <i>Single Root IOV Specification</i> . In MR, VFs exist within Virtual Hierarchies (VHs).
Virtual Hierarchy (VH)	Portion of an MR Topology assigned to a single PCIe Domain Hierarchy.
Virtual Hierarchy Number (VHN)	Link Local Number designating a VH.
Virtual Intermediary (VI)	A software component supporting one or more SIs – colloquially known as a Hypervisor or Virtual Machine Monitor. Specification of the behavior and architecture of VI is outside the scope of this specification.
Virtual Switch (VS)	A logical PCIe Switch associated with a single VH implemented in a MR Switch.





## 1. Architectural Overview

Within the industry, significant effort has been expended to increase the effective hardware resource utilization through the use of virtualization technology. The Multi-Root I/O Virtualization (MR-IOV) specification defines the extensions to the PCI Express (PCIe) specification suite to enable multiple non-coherent Root Complexes (RCs) to share PCI hardware resources.

To illustrate how this technology can be used to increase effective resource utilization, consider the following generic server blade configuration as illustrated in the Figure 1-1 below.

- ❑ The server blade configuration contains four server blades and two external fabric switches. In a high availability configuration, nominally there would be two external fabric switches of each type to avoid any single point of failure (SPOF) for a total of four switches.
- ❑ In this example, each switch provides two external connectivity ports though more can be configured to deliver high availability solutions as well as increased aggregate performance.
- ❑ Each server blade is provisioned with two PCIe Endpoint Devices – an Ethernet and a storage area network (SAN) device. This translates to a total of eight PCIe Endpoint Devices. These PCIe Devices are point-to-point connected to a Root Port (RP) [not shown] – either emitted by a chipset or a processor.
- ❑ Each I/O device and switch port is typically provisioned to enable any I/O device to operate at full bandwidth.

Depending upon workload, the example configuration's I/O resource capacity may be excessive resulting in under-utilized hardware.

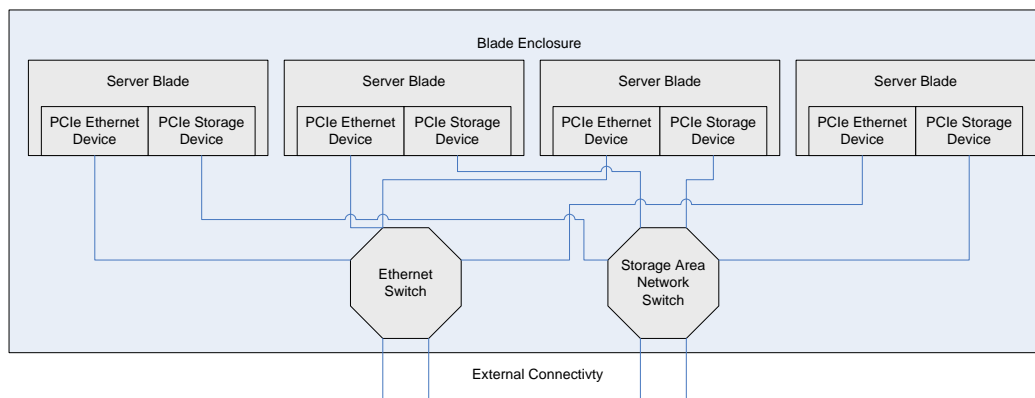
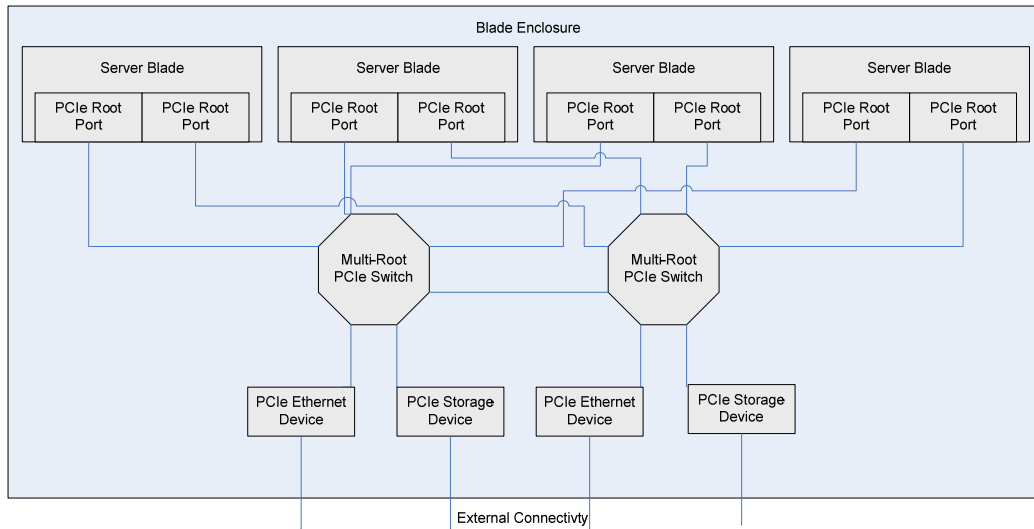


Figure 1-1: Generic Server Blade Configuration

Through the application of MR-IOV technology, the prior example server blade configuration can be transformed as illustrated in Figure 1-2.



**Figure 1-2: Example Server Blade Configuration using MR-IOV Technology**

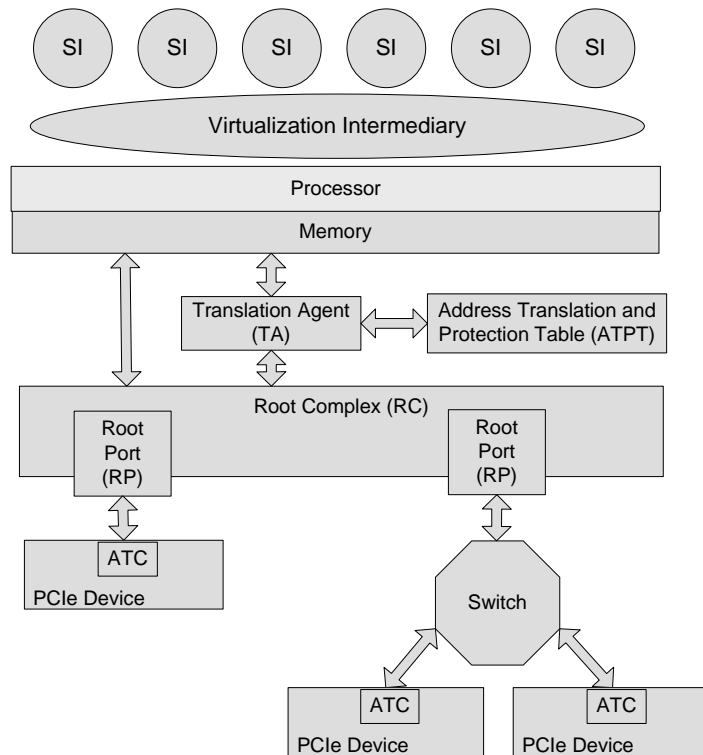
In contrast to the figure 1-1, the following can be observed:

- ❑ The server blade configuration contains four server blades. The server blades do not contain PCIe Endpoint Devices but instead connect a Root Port (RP) to a Multi-Root Aware (MRA) PCIe Switch.
- ❑ The two external fabric switches are replaced by two MRA PCIe Switches. While the details will be described in a subsequent section, the following should be noted:
  - Unlike a PCIe Switch which contains a single upstream Port and can only be claimed by a single RP, a MRA PCIe Switch contains multiple upstream Ports to enable it to connect to multiple RPs. This enables the MRA PCIe Switch to be a shared component within the configuration.
  - Multiple MRA PCIe Switches can be interconnected in a variety of topologies to create high availability solutions as well as provide increased I/O fan-out capacity.
- ❑ In place of eight PCIe Endpoint Devices – four of each type – the example MR-IOV configuration contains four MRA PCIe Endpoint Devices – two of each type. Each MRA PCIe Endpoint Device is attached to a MRA PCIe Switch downstream Port enabling each to be accessed, and thus shared, by any of the server blades.
- ❑ Unlike the prior example configuration where I/O is dedicated to each server blade, a MR-IOV based configuration enables the I/O to be dynamically assigned. A fraction or an entire I/O Device can be assigned to each server blade based on its workload requirements.

As noted above, a MR-IOV configuration reduces the component count and change the component composition. This specification will cover the elements involved in delivering a MR-IOV configuration.

## 1.1. How does MR-IOV Work?

To understand how MR-IOV works, first examine an example platform configuration devoid of any Single Root IOV (SR-IOV) – see the *Single-Root I/O Virtualization Specification* - or MR-IOV technology as illustrated in Figure 1-3.



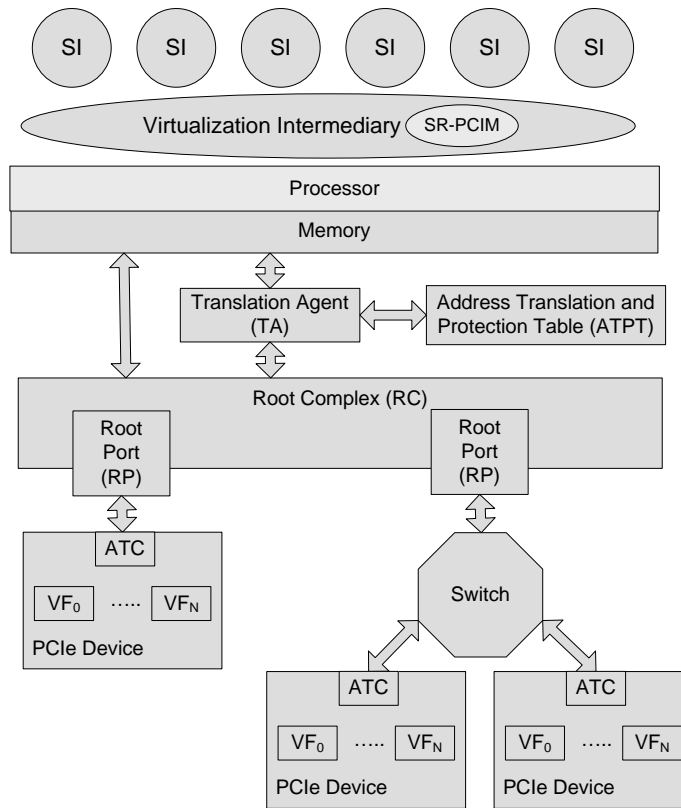
**Figure 1-3: Example Platform Configuration without SR-IOV or MR-IOV Technology**

The above example platform is composed of the following:

- ❑ A processor capable of running any operating system (OS). In virtualized environments, a processor can execute a Virtual Intermediary (VI) which abstracts or virtualizes all hardware from one or more System Images (SI). Each SI contains a virtual I/O device driver while the VI will contain the I/O-specific device drivers and perform all I/O hardware accesses.
- ❑ A memory controller and associated memory.
- ❑ A Translation Agent (TA) and Address Translation and Protection Table (ATPT).

- A TA parses the contents of a PCIe DMA request transaction (TLP) to index an ATPT to derive the physical address translation and access rights. The purposes for having DMA address translation vary and include:
  - ◆ Limiting the destructiveness of a ‘broken’ or miss-programmed DMA I/O Function
  - ◆ Providing for scatter/gather
  - ◆ Ability to redirect message-signaled interrupts (e.g., MSI or MSI-X) to different address ranges without requiring coordination with the underlying I/O Function
  - ◆ Address space conversion (32-bit I/O Function to larger system address space)
  - ◆ Virtualization support
- A PCIe Endpoint may contain an Address Translation Cache (ATC) in support of the PCI-SIG *Address Translation Services Specification*.
- A PCIe Root Complex (RC) containing one or more Root Ports (RP) with direct-attached or PCIe Switch-attached PCIe Devices or PCI / PCI-X Bridges. Each RP defines a unique hierarchy domain (see *PCI Express Base Specification*).

Now examine a platform that supports SR-IOV technology as illustrated in Figure 1-4 below.



**Figure 1-4: Example Platform Configuration with SR-IOV Technology**

The differences between the platforms in Figure 1-3 and Figure 1-4 are:

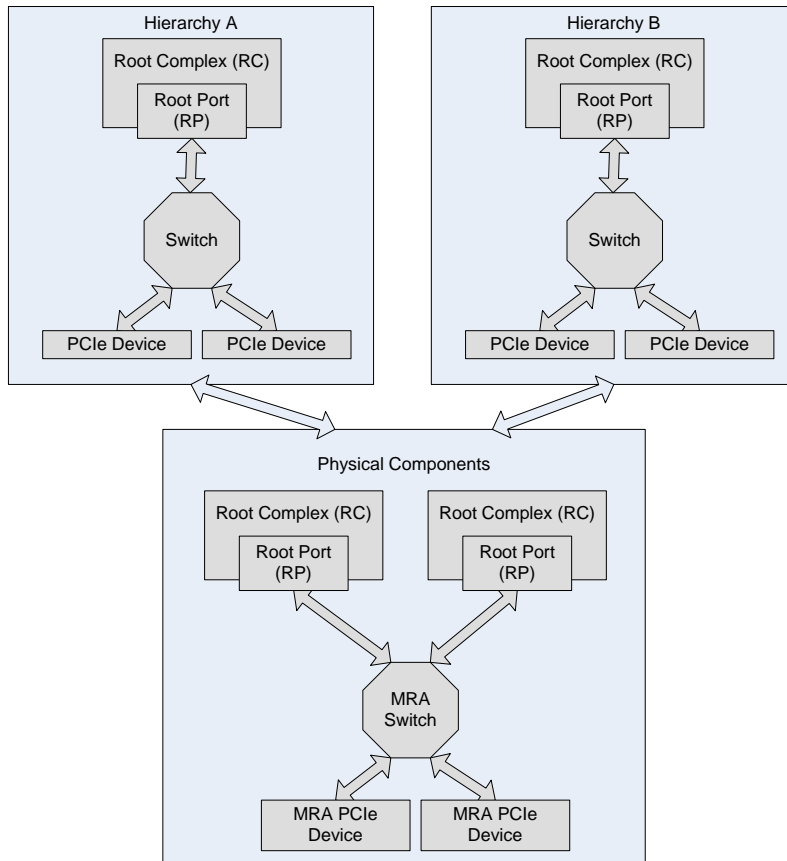
- ❑ The PCIe Device support the SR-IOV capability as defined in the *Single Root I/O Virtualization Specification*. SR-IOV enables a PCIe Device to support multiple Virtual Functions (VFs).
- ❑ SR-PCIM (Single Root PCI Manager) is responsible for SR-IOV hardware configuration and management. SR-PCIM is nominally integrated within a VI, though a variety of implementation options exist and remain outside the scope of the PCI-SIG specifications.

From the prior two figures, the following key semantics should be kept in mind:

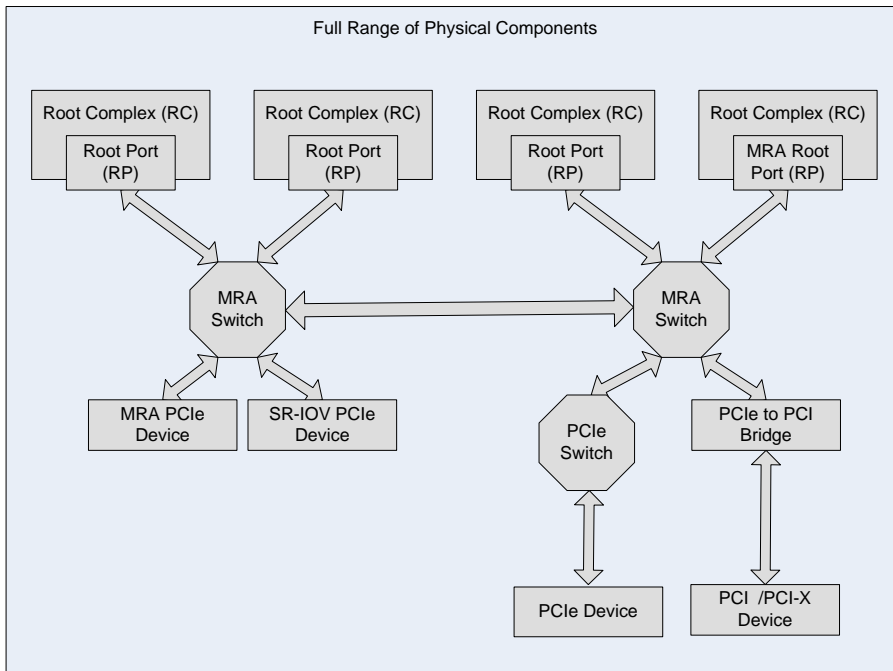
1. The OS or VI has exclusive control over each PCIe Component – RC, RP, Switch, Link, PCIe-to-PCI/PCI-X Bridge, Device, and Function.
  - a. All PCI enumeration, configuration operations, reset, power management, and event handling, e.g. errors, must only be initiated or processed by the OS or VI. Operations initiated by a SI are trapped by the VI and processed on its behalf.
2. Each RP acts as the terminus for all upstream targeted operations, e.g. error event notification.

In order to support either example platform while preserving these semantics, the PCI components underneath each RP must be virtualized and logically overlaid on the MRA PCIe Switches and Devices as illustrated in Figure 1-5. The virtualized PCI components are referred to as a Virtual Hierarchy (VH). A VH has the following attributes:

- ❑ Each VH must contain at least one PCIe Switch.
  - The PCIe Switch will be a virtualized component implemented over of a MRA Switch.
  - The PCIe Switch functionality and semantics are per the *PCI Express Base Specification*.
- ❑ Each VH may contain any mix of PCIe Devices, MRA PCIe Devices, or PCIe to PCI / PCI-X Bridges as illustrated in Figure 1-6.
  - A PCIe Device is a device that does not support the MR-IOV Capability. Such a device must only be visible in a single VH at a time. A PCIe Device can be serially shared among a set of accessible VH within the MR-IOV topology. The PCIe Device is serially deleted from the current source VH and added to destination VH.
  - A PCIe Device may support the SR-IOV Capability which enables it to be shared by multiple SI executing above a single RP.
  - A PCIe to PCI / PCI-X Bridge can only be visible in a single VH at a time. As with a PCIe Device, a PCIe to PCI / PCI-X Bridge can be serially shared among a set of accessible VH within the MR-IOV topology using a conceptually similar deletion / addition process as a PCIe Device.
    - ◆ The SR-IOV Capability does not apply to the PCIe to PCI / PCI-X Bridge. The bridge and all associated PCI / PCI-X devices can only be configured in a single OS, VI, or SI at a time.
  - A MRA PCIe Device is a device which supports the MR-IOV Capability. Such a device can be visible in multiple VH at a time depending upon the MR-IOV resources provisioned. A MRA PCIe Device can be added or deleted from any accessible VH within a MR-IOV topology.
  - A MRA PCIe Device must support the SR-IOV Capability. This enables it to be shared by multiple SI executing above each RP.
- ❑ The MR-IOV topology must contain at least one MRA PCIe Switch.
  - Multiple MRA PCIe Switches can be provisioned and interconnected in a variety of topologies – tree, fat-tree, star, mesh, etc.
  - A MRA PCIe Switch must contain two or more upstream Ports. Each upstream Port must connect to a RP which acts as the root of the VH.



**Figure 1-5: Two Virtual Hierarchies (VH) Implemented over Shared Physical Components**



**Figure 1-6: Physical Components that can be supported in a MR-IOV Topology**

### 1.1.1. MRA Components

The prior section illustrated that MR-IOV is primarily the overlaying of multiple VH over a shared physical set of MRA and non-MRA components. To further understand how MR-IOV works, let's examine the MRA components in more detail.

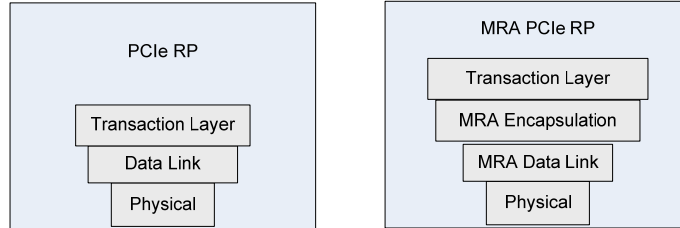
#### 1.1.1.1. Multi-Root Aware Root Port (MRA RP)

As illustrated in Figure 1-6, a PCIe RP supporting either a single OS or a VI with multiple SI can be connected to a MRA Switch and access multiple downstream devices and bridges. The PCIe RP though, is restricted to a single VH. In order to enable multiple VH to be accessed, a MRA PCIe RP is required. A MRA PCIe RP differs from a PCIe RP in the following ways:

- ❑ A MRA PCIe RP maintains state to delineate each VH. At a high level, this amounts to a set of resource mapping tables to translate the I/O function associated with each SI into a VH and MR I/O function identifier.
- ❑ A MRA PCIe RP participates in the MR transaction encapsulation protocol (see subsequent section for details) to enable a MRA PCIe Switch to derive the VH and associated routing information.



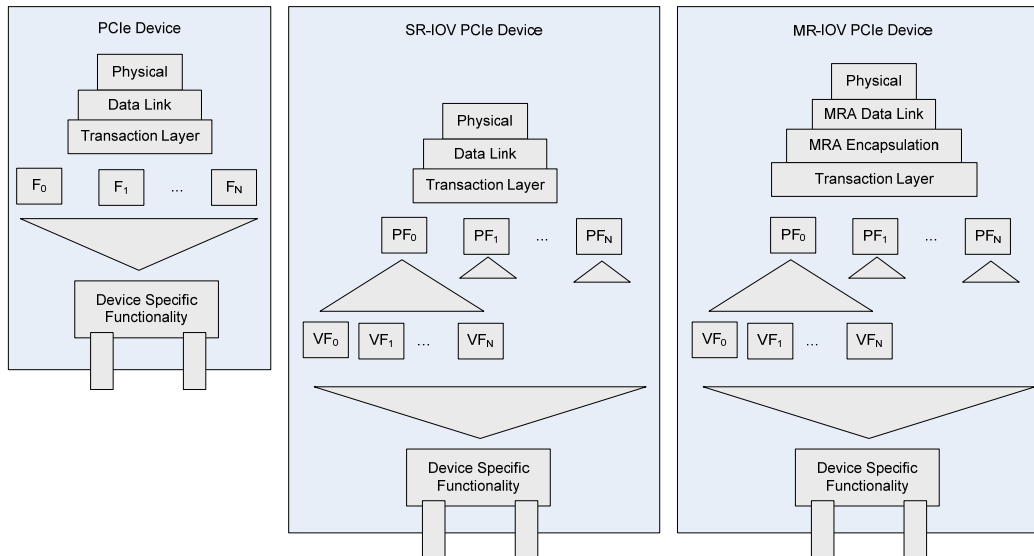
- ❑ A MRA PCIe RP emits a MRA Link. A MRA Link is identical the physical layer to a PCIe Link as defined in the *PCI Express Base Specification*. A MRA Link differs at the data link layer where a new set of DLLP are defined to support the MR-IOV protocol.
- ❑ A MRA PCIe RP may implement MRA congestion management (see subsequent chapter for details).



**Figure 1-7 PCIe RP and MRA PCIe RP Functional Block Comparison**

#### 1.1.1.2. Multi-Root Aware PCIe Device (MRA PCIe Device)

A MR-IOV platform may contain any mix of PCIe Devices, SR-IOV PCIe Devices, or MR-IOV PCIe Devices. Figure 1-8 illustrates a functional block comparison between these three types of devices.



**Figure 1-8 PCIe Device, SR-IOV, and MRA PCIe Device Functional Block Comparison**

A MRA IOV PCIe Device differs from a PCIe Device and a SR-IOV PCIe Device in the following ways:

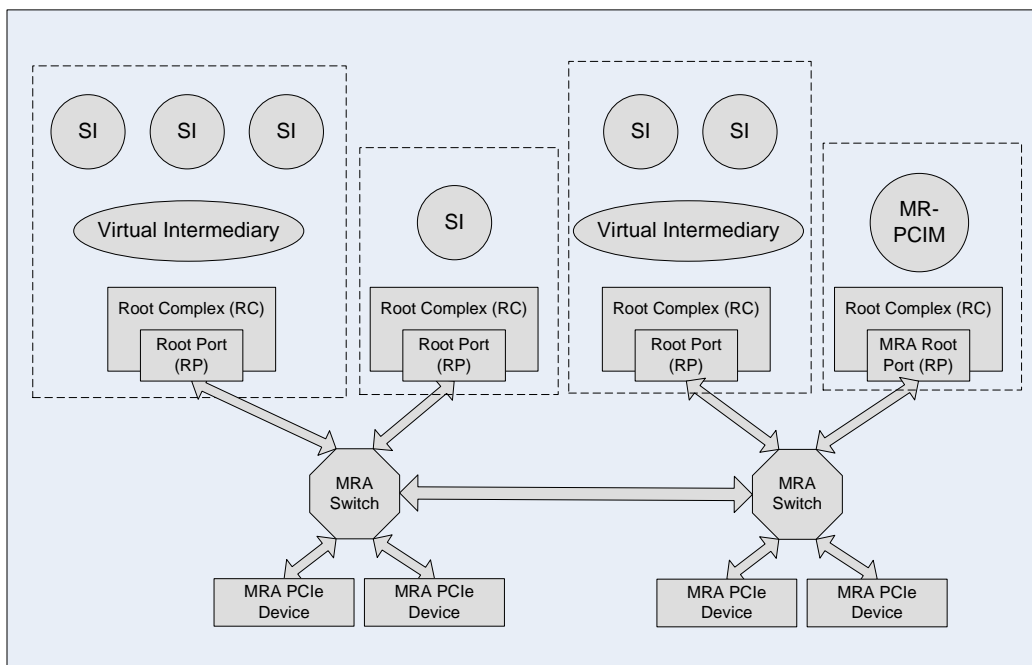
- ❑ The MRA IOV PCIe Device must support the new MR DLLP protocol.
  - A PCIe Device or a SR-IOV PCIe Device do not support the MR-IOV capability and therefore are unable to participate in this protocol. A MRA PCIe Switch must subsume all responsibility for forwarding transactions and event handling on behalf of these devices through the MR-IOV topology. The MRA PCIe Switch will perform all encapsulation or de-encapsulation as appropriate.
- ❑ The MRA IOV PCIe Device must support the MR-IOV transaction encapsulation protocol.
  - The MR-IOV encapsulation protocol provides VH identification information to the MRA PCIe Switch to enable the transaction to be transparently forwarded through the MR-IOV topology without requiring modification to the *PCI Express Base Specification* TLP protocol or contents.
- ❑ The MRA IOC PCIe Device is composed of a set of Physical Functions (PF).
  - Each PF supports a full PCI Configuration and PCIe Extended Configuration Space.
  - Each PF supports a full BAR.
  - Each PF must only be assigned to a single accessible VH at a given time.
  - One or more PF may be assigned to any accessible VH.
  - A PF and its associated resources may be migrated from one VH to another.
  - Each PF may support zero or more Virtual Functions (VF).
    - ◆ VF share resources including some portions of the configuration space with the associated PF.
    - ◆ A VF exists only within the VH associated with the PF.
    - ◆ A MRA PCIe Device with multiple PF and zero VF per PF is conceptually equivalent to a single function PCIe Device configured per VH.
  - The number of VF provisioned per PF may vary on a per PF basis.
  - Each PF represents a single device-specific functionality, e.g. an Ethernet controller, a SATA controller, etc. Subsequently, each VF must represent the same device-specific functionality. This enables the existing device driver models to be supported.

#### 1.1.1.3. Multi-Root PCI Manager (MR-PCIM)

Each MRA component must support a corresponding MR-IOV capability. This capability is accessed and configured by the Multi Root PCI Manager (MR-PCIM). MR-PCIM can be implemented anywhere within the MR-IOV topology – for example, above a RP as illustrated in Figure 1-9, or, for example, through a private interface provided by a MRA PCIe Switch.

MR-PCIM is responsibilities include:

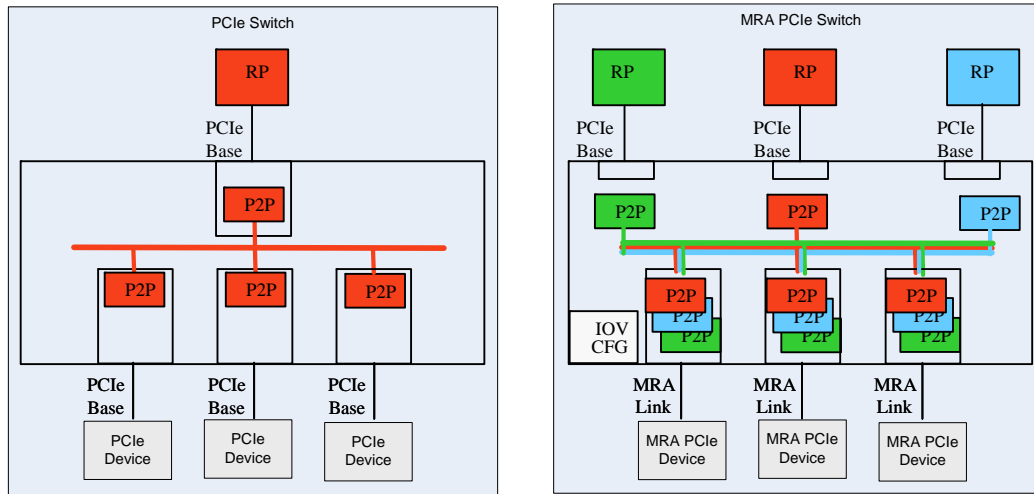
- ❑ Enumeration of the physical components within the MR-IOV topology. MR-PCIM must determine what components are or are not MR-IOV capable, how components are interconnected, and what the PCIe and MR-IOV resources they provide.
- ❑ MR-PCIM configures the components and resources that comprise each VH. The policies to determine this are outside the scope of this specification.
- ❑ Given the physical hardware is shared among a set of VH, MR-PCIM configures all PCIe and MR-IOV attributes including: link signaling rate, VC arbitration, Port arbitration, Access Control Services (ACS), etc.
- ❑ Given the physical hardware is shared among a set of VH, MR-PCIM processes or controls various events, e.g. RESET, physical hardware failure, surprise add / remove, error handling, etc.
- ❑ <continue to build up list of responsibilities>



**Figure 1-9 MRA PCIM in a MR-IOV Topology**

#### *1.1.1.4. Multi-Root Aware PCIe Switch (MRA PCIe Switch)*

In a prior section it was noted that a MRA Switch is conceptually the overlay of multiple PCIe Switches onto a single physical package. This is illustrated in more detail in Figure 1-9.



**Figure 1-9 PCIe Switch and MRA PCIe Switch Functional Block Comparison**

The PCIe Switch is composed of a set of logical P2P bridges with a single upstream Port attached to a PCIe RP and one or more downstream Ports attached to either a PCIe Device or a PCIe to PCI / PCI-X Bridge. A PCIe Switch also operates using a single address space.

In contrast a PCIe Switch, a MRA Switch is as follows:

- ❑ A MRA Switch is composed of one or more upstream Ports attached to either PCIe RP or MRA PCIe RP or the downstream Port of a MRA Switch.
  - If the upstream Port is attached to a PCIe RP, the MRA Switch must transparently provide all MRA related services on behalf of the PCIe RP.
- ❑ A MRA Switch is composed of one or more downstream Ports attached to PCIe Devices, MRA PCIe Devices, PCIe Switch upstream Ports, MRA Switch upstream Ports, or PCIe to PCI / PCI-X Bridges.
- ❑ A set of logical P2P bridges that constitute a VH. A MRA Switch must support two or more VH.
- ❑ Each VH represent a separate address space. The combination of a VH identifier and the address contained within the PCIe TLP enable the MRA Switch to forward the a TLP to appropriate egress Port as well as a MRA RP or MRA PCIe Device to delineate which PF or VF is the source or sink of the PCIe TLP.
- ❑ <fill in additional attributes / operational semantics to flesh out this section>

### 1.1.2. MR Initialization Overview

<td>

### 1.1.3. MR Transaction Encapsulation Overview

<likely migrate sections within chapter 2 here>

### 1.1.4. MR Congestion Management Overview

<td>

### 1.1.5. MR Error and Event Handling Overview

<td>

### 1.1.6. MR-IOV and ARI (Alternative Routing Identifier)

<td>

### 1.1.7. MR-IOV Relationship to SR-IOV and ATS

Briefly:

- ❑ A MRA PCIe Device must support the SR-IOV Capability per the *Single Root I/O Virtualization Specification*.
  - A SR-IOV PCIe Device must support the PCI Express Base Specification.
  - By requiring these specifications to be supported, the number of permutations is reduced further enhancing the ability deploy and interoperate across a wide range of solution options.
- ❑ A MRA PCIe Device may support the Address Translation Services Specification.

## 1.2. Overview of MR Transaction Layer

Figure 1-2 shows an example Multi-Root Topology

Deleted: Figure 1-2

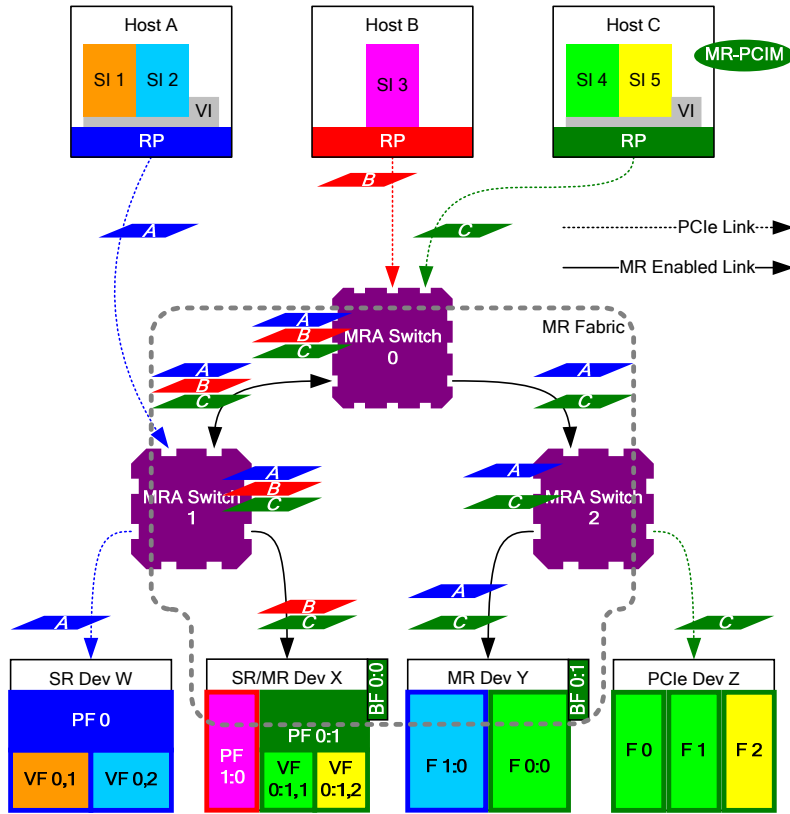


Figure 1-2: Example Multi-Root Topology

The solid links are using the Multi-Root Wire Protocol described in this specification. The dashed links are using the PCIe Protocol. Functions on SR Devices are designated **PF *f*** and **VF *f:s*** where *f* represents the Function Number of the PF and *s* indicates which VF Slot belonging to the PF is involved. Functions on MR Devices are designated **PF *h:f***, **VF *h:f:s*** or **F *h:f*** where *h* is added to indicate which Virtual Hierarchy (VH) is involved.

This example shows a single Multi-Root Topology. TLPs inside the MR Topology are labeled with the VH they belong to. TLPs outside the MR Topology belong to a single VH and no label is needed. The MR Ingress point is the point where a TLP first encounters an MR Topology. The MR Egress point is where a TLP Exits an MR Topology. These points exist inside some MR component (in this example, the Switches and Devices X and Y).

In this example, all Root Ports use PCIe protocol. Each Root Port is the root of a PCIe Hierarchy. There are three VHs (A, B, C) each associated with one of the root Ports. Hosts A and C are running a Virtual Intermediary that supports SR sharing. B is running an Operating System directly (no Virtual Intermediary is involved).

In this example, there are four Devices, one of each variety.

- ❑ Device W is a Single-Root IOV Device. It is assigned to VH A. Two System Images (SIs) are in use on Host A. The Virtual Intermediary on Host A has further assigned VF 0,1 to SI 1 and VF 0,2 to SI 2.
- ❑ Device X is using both Single Root and Multi-Root sharing. PF 1:0 is assigned to VH B. PF 0:1 is assigned to VH C. The Virtual Intermediary running on Host C has further assigned VF 0:1,1 to SI 4 and VF 0:1,2 to SI 5. The MR features of the device are managed through the Base Function which is assigned to VH C.
- ❑ Device Y is using only Multi-Root sharing. F 1:0 is assigned to VH A and F 0:0 is assigned to VH C. In VH A, the Virtual Intermediary has further assigned F 1:0 to SI 2. In VH C, the Virtual Intermediary has further assigned F 0:0 to SI 4. The MR features of the device are managed through the Base Function which is assigned to VH C.
- ❑ Device Z is a 3 Function PCIe Device. It is assigned to VH C. Virtual Intermediary software on VH C has further assigned F 0 and F 1 to SI 4 and F 2 to SI 5.

All Switches shown in this example are Multi-Root Aware (MRA). Non-MRA Switches are also possible; however, such Switches and all components below them will be associated with a single Root Port in a single VH. Note that this non-MR sub-tree can be a mixture of SR aware and non-SR aware PCIe components.

Multi-Root Aware Components enforce separation between VHs. Software running in one VH is not allowed to affect other VHs. For example, every VH has a complete and independent address space.

[Figure 1-3](#), shows the same example as [Figure 1-2](#), but only shows the components visible to Host A. The MRA Switches and Devices appear to software as Single Root equivalents.

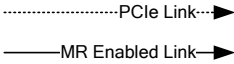
Similarly, [Figure 1-4](#) also shows the example from [Figure 1-2](#) but shows only components visible to Host C. Note that the link between Switch 0 and Switch 1 changes direction between these views of the topology. In Multi-Root systems, the logical upstream / downstream direction of a link is a per-VH concept and is distinct from physical link direction that was established during link bring up.

Deleted: Figure 1-3

Deleted: Figure 1-2

Deleted: Figure 1-4

Deleted: Figure 1-2



**Figure 1-3: Example Multi-Root Topology as viewed from Host A**



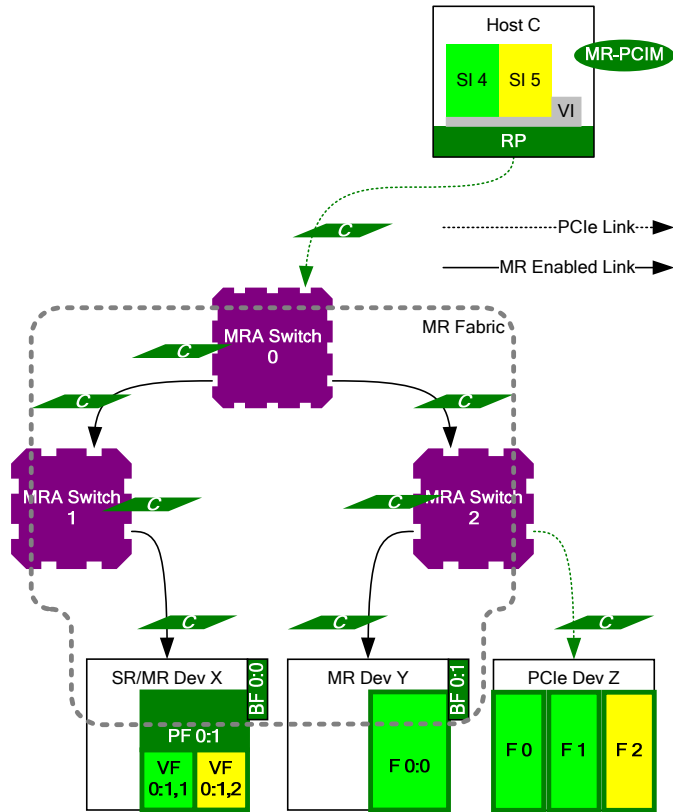


Figure 1-4: Example Multi-Root Topology as viewed from Host C



## 2. MR Protocol Changes

There are five parts of the PCIe Protocol that are changed to support Multi-Root operation.

- ❑ Negotiating use of the MR link protocol
- ❑ Tagging TLPs with a TLP Prefix Header
- ❑ Supporting per-VH equivalent of Hot Reset
- ❑ Supporting enhanced, per-VH flow control
- ❑ Processing of certain messages (e.g. INTx, PME)

These will be discussed in the following sections.

### 2.1. MR Link Protocol Negotiation

MR Links use an enhanced link protocol. For each link, the use of this enhanced link protocol is determined by a negotiation between link partners. This negotiation occurs after the Physical Layer's Link training but before PCIe Flow Control negotiation.

During this negotiation, MR components determine whether their Link Partner agrees to use the MR link protocol and which version of the link protocol to use. They also communicate certain MR parameters (MaxVH and MaxVL).

This negotiation occurs by using the new MRInit DLLP as shown in [Figure 2-1](#) and [Table 2-1](#).

Deleted: Figure 2-1

Deleted: Table 2-1

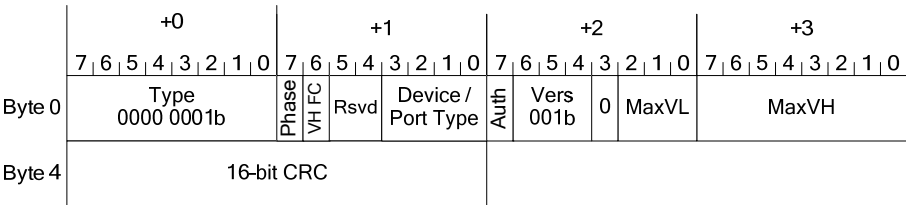


Figure 2-1: MRInit DLLP Format

**Table 2-1: MRInit DLLP Fields**

<b>Location</b>	<b>Description</b>
Byte 0 Bits 7:0	<b>Type</b> – The value 0000 0001b indicates an MRInit DLLP.
Byte 1 Bit 7	<b>Phase</b> – Indicates the phase of the MR Negotiation protocol.
Byte 1 Bit 6	<b>VH FC</b> – If Set, indicates that the sender supports per-VH, VL flow control. If Clear, indicates that the sender only supports per-VL flow control. Must be set for Switches.
Byte 1 Bits 5:4	<b>Reserved</b>
Byte 1 Bits 3:0	<b>Device / Port Type</b> – Device / Port Type of the sender. Encoding is identical to the Device / Port Type field in the PCI Express Capability (Offset 02, Bits 7:4).
Byte 2 Bit 7	<b>Authorized</b> – If Device / Port Type indicates a Switch, indicates that the sender is an Authorized port on an MR Capable Switch. Must be 0b if Device / Port Type does not indicate a Switch.
Byte 2 Bits 6:4	<b>Protocol Version</b> – must be 001b for this version of the specification.
Byte 2 Bit 3	<b>Reserved</b> – Transmit 0b.
Byte 2 Bits 2:0	<b>MaxVL</b> – Maximum number of Virtual Links supported by the sender.
Byte 3	<b>MaxVH</b> – Maximum number of Virtual Hierarchies supported by the sender.

The MRInit DLLP is a new encoding, not defined in PCI Express. PCI Express components are required to ignore DLLPs not defined in the Base PCI Express Specification (see section 3.5.2.2 of the PCI Express 2.0 Specification, or section 3.5.2.1 of the PCI Express 1.1 Specification).

MR Devices will always negotiate to use the MR link protocol. MRA Switches and Root Ports will only negotiate if enabled to do so.

The negotiation sequence between two MR components that are enabled to use the link in MR mode is shown in [Figure 2-2](#).

**Deleted:** Figure 2-2

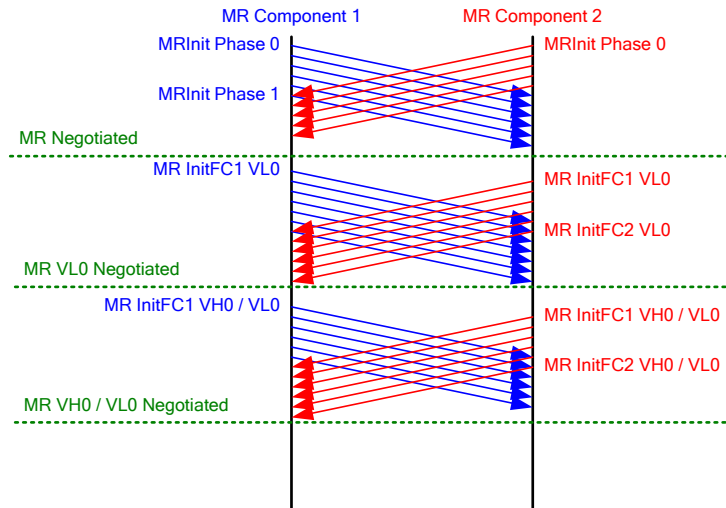


Figure 2-2: MR to MR Initialization Sequence

The negotiation sequence between an MR component and a Base component (or an MR component that is not enabled to use MR on the link) is shown in [Figure 2-3](#).

Deleted: Figure 2-3

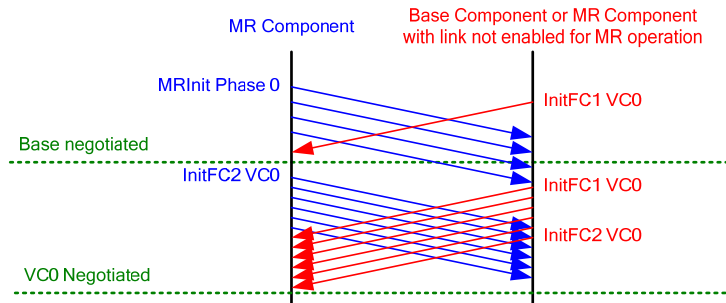


Figure 2-3: MR to Base Initialization

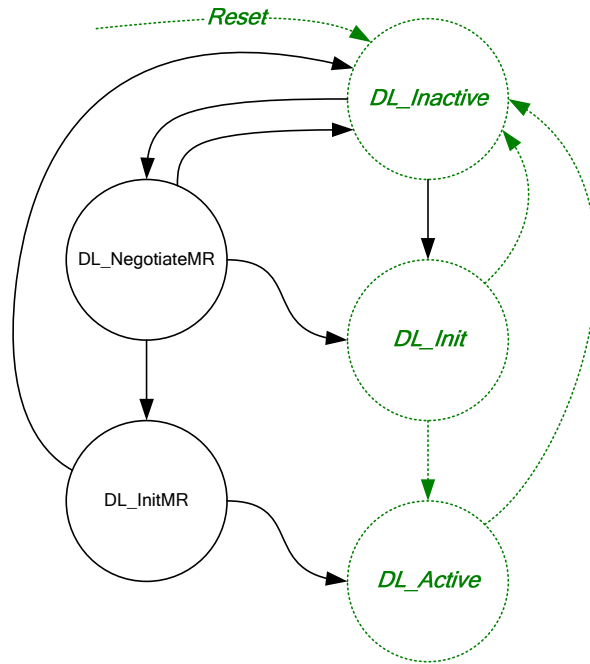
The Data Link Control and Management State Machine (DLCMSM) is modified for MR operation. The states for this machine are described below, and are shown in [Figure 2-4](#). Italic text and green dashed lines represent Base PCI Express.

Deleted: Figure 2-4

States:

- ❑ *DL\_Inactive* – Physical Layer reporting Link is non-operational or nothing is connected to the Port
- ❑ *DL\_Init* – Physical Layer reporting Link is operational, initialize Base Flow Control for the default Virtual Channel

- ❑ DL\_NegotiateMR – Physical Layer reporting Link is operational, negotiate the use of the MR Link Protocol
- ❑ DL\_InitMR – Physical Link reporting Link is operational, initialize MR Flow Control for VL0 and for VH0 / VL0
- ❑ *DL\_Active – Normal operation mode*



**Figure 2-4: MR Data Link Control and Management State Machine (MR-DLCMSM)**

The *DL\_Inactive* state rules are modified as follows:

- ❑ *DL\_Inactive*

...

- *Exit to DL\_Init if:*
  - ◆ *Indication from the Transaction Layer that the Link is not disabled by software, the link is not enabled for MR operation and the Physical Layer reports Physical LinkUp = 1b*
- *Exit to DL\_NegotiateMR if:*
  - ◆ *Indication from the Transaction Layer that the Link is not disabled by software, the link is enabled for MR operation and the Physical Layer reports Physical LinkUp = 1b*

Rules are added for the new DL\_NegotiateMR and DL\_InitMR states:

❑ DL\_NegotiateMR

- While in DL\_NegotiateMR:
  - ◆ Negotiate MR Link Protocol usage following the MR Link Protocol Negotiation described in Section 2.1.1
  - ◆ Report DL\_Down status
  - ◆ The Data Link Layer of a Port with DL\_Down status is permitted to discard any received TLPs provided that it does not acknowledge those TLPs by sending one or more Ack DLLPs
- Exit to DL\_Init if:
  - ◆ MR Link Protocol negotiation completes indicating PCIe Link Mode and the Physical Layer continues to report Physical LinkUp = 1b
- Exit to DL\_InitMR if:
  - ◆ MR Link Protocol negotiation completes indicating MR Link Mode and the Physical Layer continues to report Physical LinkUp = 1b
- Terminate attempt to negotiate MR Link Protocol and Exit to DL\_Inactive if:
  - ◆ Physical Layer reports Physical LinkUp = 0b

❑ DL\_InitMR

- While in DL\_InitMR:
  - ◆ Initialize Flow Control for the default Virtual Link, VL0, and default Virtual Hierarchy on the default Virtual Link, VH0/VL0, following the Flow Control initialization protocol described in Section 2.1.2
  - ◆ Report DL\_Down status while in state MRFC\_INIT1\_VL, MRFC\_INIT2\_VL or MRFC\_INIT1\_VH; DL\_Up status in state MRFC\_INIT2\_VH
  - ◆ The Data Link Layer of a Port with DL\_Down status is permitted to discard any received TLPs provided that it does not acknowledge those TLPs by sending one or more Ack DLLPs
- Exit to DL\_Active if:
  - ◆ Flow Control initialization completes successfully, and the Physical Layer continues to report Physical LinkUp = 1b
- Terminate attempt to initialize Flow Control and Exit to DL\_Inactive if:
  - ◆ Physical Layer reports Physical LinkUp = 0b

### 2.1.1. MR Link Protocol Negotiation

The MR Link Protocol Negotiation involves two phases.

- ❑ Phase 0 entered when MR Link Protocol Negotiation is required
  - Entrance to DL\_NegotiateMR state
- ❑ While in Phase 0 of the MR Link Protocol Negotiation:
  - Transaction Layer must block transmission of TLPs and DLLPs other than the MRInit DLLP
  - Continuously transmit an MRInit DLLP as shown in [Figure 2-1](#), MaxVL, MaxVH, Auth and Device / Port Type reflect the sender's values. Protocol Version is 1h. Phase is 0b.
    - ◆ This does not block Physical Layer initiated transmissions (for example, Ordered Sets)
  - Process received MRInit DLLPs:
    - ◆ Record the MaxVL, MaxVH, VH FC, Device / Port Type and Authorized values
  - Exit to Phase 1 of the MR Link Protocol Negotiation if:
    - ◆ MRInit DLLP was received with Protocol Version 1h (with either Phase).
  - Exit indicating PCIe Link Mode if:
    - ◆ InitFC1 DLLP was received.
- ❑ While in Phase 1 of the MR Link Protocol:
  - Transaction Layer must block transmission of TLPs and DLLPs other than the MRInit DLLP
  - Continuously transmit an MRInit DLLP as shown in [Figure 2-1](#), MaxVL, MaxVH, Auth and Device / Port Type reflect the sender's values. Protocol Version is 1h. Phase is 1b.
    - ◆ This does not block Physical Layer initiated transmissions (for example, Ordered Sets)
  - Process received MRInit DLLPs:
    - ◆ Ignore the MaxVL, MaxVH, VF FC, Device / Port Type and Authorized values
  - Exit indicating MR Link Mode if either:
    - ◆ MRInit DLLP was received with Protocol Version 1h and Phase 1b
    - ◆ Any MRInitFC1\_VL DLLP was received

Deleted: Figure 2-1

Deleted: Figure 2-1

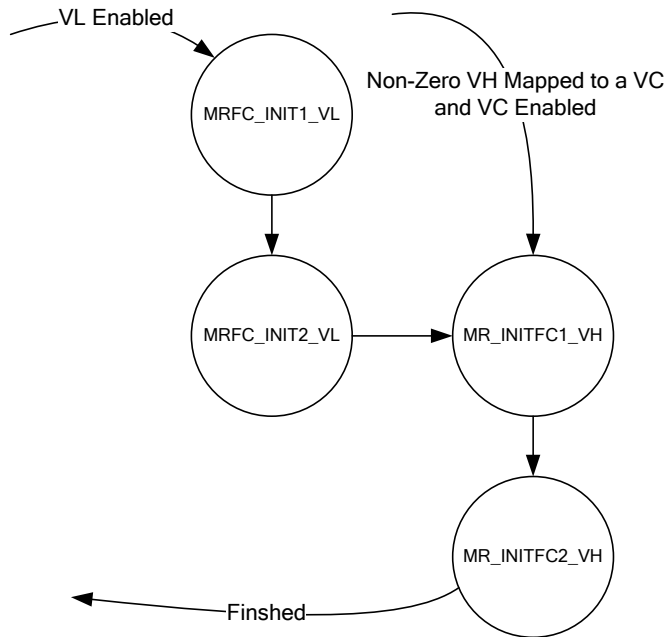
### 2.1.2. MR Flow Control Initialization Protocol

Before starting normal operation following power-up or interconnect Reset, it is necessary to initialize Flow Control for the default Virtual Link, VL0 and the default Virtual Hierarchy, VH0. In addition, when additional Virtual Links (VLs) and Virtual Hierarchies (VHs) are enabled, the Flow Control initialization process must be completed for each newly enabled VL or VH before it can be used. This section describes the initialization process that is used for all VLs and all VHs. Note that since VL0 and VH0 are enabled before all other VLs and VHs, no TLP traffic of any kind will be

active prior to initialization of VL0 and VH0. However, when additional VLs or VHs are being initialized there will typically be TLP traffic flowing on other, already enabled, VLs and VHs. Such traffic has no direct effect on the initialization process for the additional VL(s) and VH(s).

There are four states in the MR VL/VH initialization process. These states are shown in [Figure 2-5](#).

Deleted: Figure 2-5



**Figure 2-5: MR InitFC State Machine**

There is a distinct instance of this state machine for each enabled VH and VL. Initialization between different enabled VLs and VHs proceeds in parallel.

State MRFC\_INIT1\_VL is entered when a VL is enabled either automatically (VH0) or explicitly (VL1-7). State MRFC\_INIT1\_VH is entered either from MR\_INITFC2\_VL or because some VH was mapped by MR-PCIM and enabled by software in the VH resulting in a VH for which flow control negotiation has not yet occurred.

The rules for this process are given in the following Section 2.1.2.2.

### 2.1.2.1. MR Flow Control DLLP Encoding

MR Flow Control DLLPs need to communicate the VH number in addition to the Base PCIe information. It is no longer possible to fit this information in a single DLLP. Consequently, for MR-IOV, Header and Data credits are communicated using different DLLPs. The formats for the various MF Flow Control DLLPs are shown in [Figure 2-6](#) through [Figure 2-15](#). The DLLP fields are

Deleted: Figure 2-6

Deleted: Figure 2-15



described in [Table 2-2](#). If a Receiver advertises infinite VH credits, then the Receiver must transmit PCIe Base UpdateFC DLLPs for that VL instead of the MRUpdateFC DLLPs described below.

Deleted: Table 2-2

- ❑ The VL Credit Type is implicitly determined from the DLLP Type Encoding used by the UpdateFC DLLP
- ❑ The VC ID field in the UpdateFC DLLP contains the VL number.
- ❑ The HdrFC field in the UpdateFC DLLP contains VL header credit value for the indicated type (P, NP, or Cpl)
- ❑ The DataFC field in the UpdateFC DLLP contains the VL data credit value for the indicated type (P, NP, or Cpl)

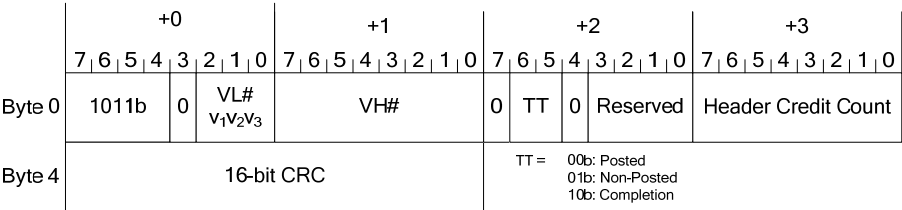
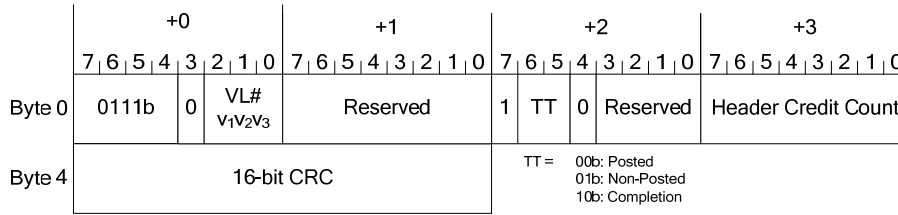
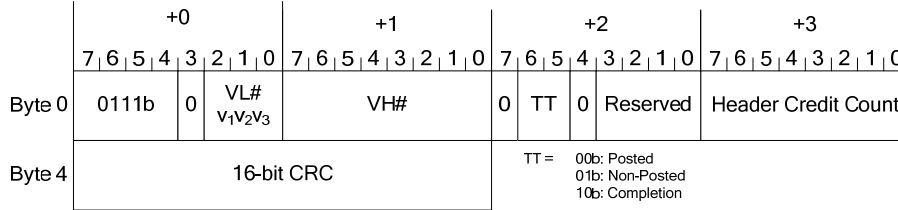


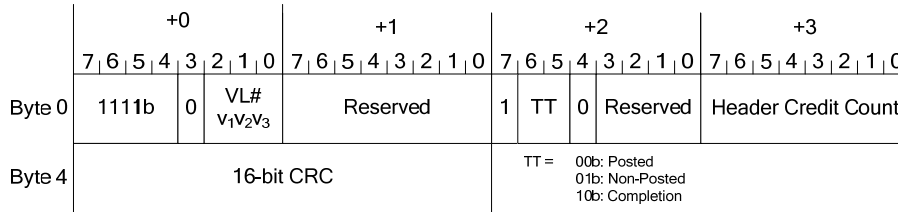
Figure 2-6: MRUpdateFC Header DLLP



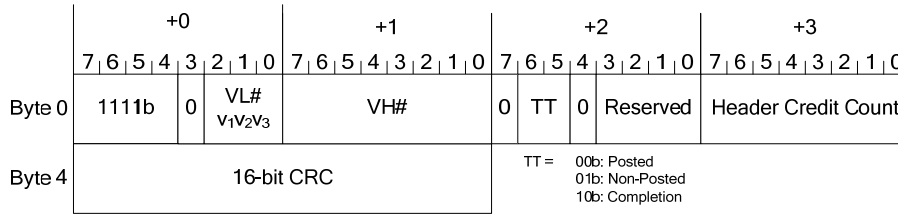
**Figure 2-7: MRInitFC1\_VL Header DLLP**



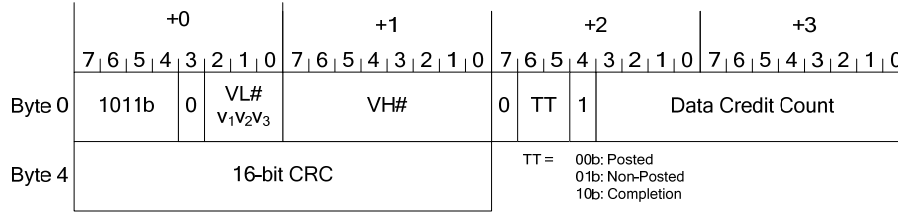
**Figure 2-8: MRInitFC1\_VH Header DLLP**



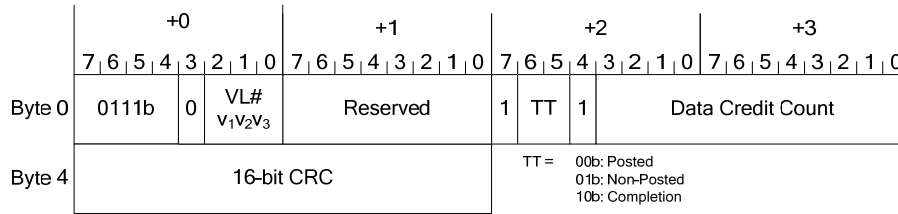
**Figure 2-9: MRInitFC2\_VL Header DLLP**



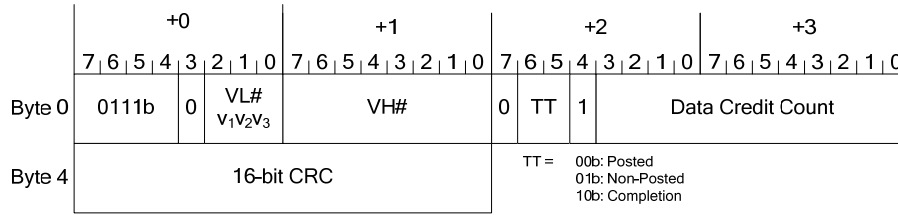
**Figure 2-10: MRInitFC2\_VH Header DLLP**



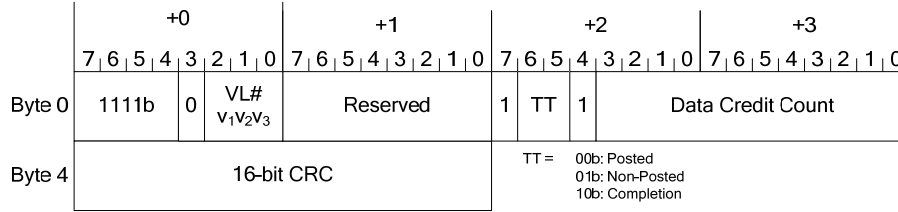
**Figure 2-11: MRUpdateFC Data DLLP**



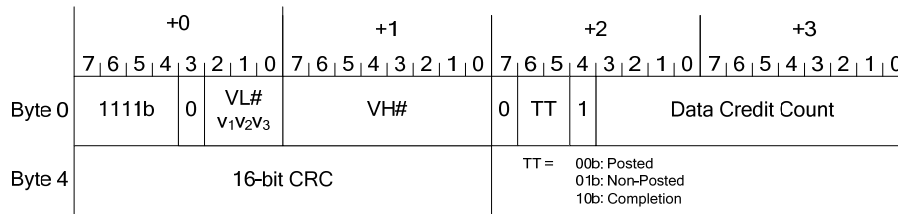
**Figure 2-12: MRInitFC1\_VL Data DLLP**



**Figure 2-13: MRInitFC1\_VH Data DLLP**



**Figure 2-14: MRInitFC1\_VL Data DLLP**



**Figure 2-15: MRInitFC2\_VH Data DLLP**

**Table 2-2: MR Flow Control DLLP Fields**

Location	Description
Byte 0 Bits 7:4	<b>DLLP Type</b> – 1011b indicates an MRUpdateFC DLLP 0111b indicates an MRInitFC1_VL or MRInitFC1_VH DLLP 1111b indicates an MRInitFC2_VL or MRInitFC2_VH DLLP
Byte 0 Bits 2:0	<b>VL Number</b> – Indicates the Virtual Link
Byte 1	<b>VH Number</b> – Indicates the Virtual Hierarchy. This field is reserved if VHO is Set.
Byte 2 Bit 7	<b>VH Omitted</b> – Indicates whether the VH Number field is present in the DLLP. If Set, this indicates the DLLP is MRInitFC1_VL or MRInitFC2_VL and the VH Number is omitted.
Byte 2 Bits 6:5	<b>TT – TLP Type</b> 00b indicates Posted credit 01b indicates Non-Posted credit 10b indicates Completion credit 11b is Reserved
Byte 3 Bit 4	<b>Credit Type</b> – 0 indicates Header Credit 1 indicates Data Credit
Byte 3 Bits 3:0 & Byte 4	<b>Data Credit Value</b> – If Credit Type is Set. PCIe encoding applies (i.e. during initialization, zero means infinite)
Byte 4	<b>Header Credit Value</b> – If Credit Type is Clear. PCIe encoding applies (i.e. during initialization, zero means infinite)

### 2.1.2.2. MR Flow Control Initialization State Machine Rules

- ❑ If at any time during initialization for VLs 1-7 a VL is disabled, any flow control initialization process involving that VL is terminated
- ❑ Rules for state MRFC\_INIT1\_VL:
  - Entered when initialization of a VL (VL<sub>x</sub>) is required
    - ◆ Entrance to DL\_InitMR state (VL<sub>x</sub> = VL<sub>0</sub>)
    - ◆ When a VL (VL<sub>x</sub> = VL<sub>1</sub>-7) is enabled by software (see Sections 4.2.1.3 and 4.3.3.2)
  - While in MRFC\_INIT1\_VL:
    - ◆ Transaction Layer must block transmission of TLPs using VL<sub>x</sub>
    - ◆ Transmit the following six MRInitFC1\_VL DLLPs for VL<sub>x</sub> in the following relative order:
      - MRInitFC1\_VL – P – Header (first)
      - MRInitFC1\_VL – P – Data (second)
      - MRInitFC1\_VL – NP – Header (third)
      - MRInitFC1\_VL – NP – Data (fourth)

- MRInitFC1\_VL – Cpl – Header (fifth)
  - MRInitFC1\_VL – Cpl – Data (sixth)
  - ◆ The six MRInitFC1 DLLPs must be transmitted at least once every 34  $\mu$ s.
    - Time spent in the Recovery LTSSM state does not contribute to this limit.
    - It is strongly encouraged that the MRInitFC1 DLLP transmissions are repeated frequently, particularly when there are no other TLPs or DLLPs available for transmission.
  - ◆ Except as needed to ensure at least the required frequency of MRInitFC1 DLLP transmission, the Data Link Layer must not block other transmissions
    - Note that this includes all Physical Layer initiated transmissions (for example, Ordered Sets), Ack and Nak DLLPs (when applicable), and TLPs using VLs and VHs that have previously completed initialization (when applicable)
  - ◆ Process received MRInitFC1\_VL and MRInitFC2\_VL DLLPs for VLx:
    - Record the indicated FC unit values
    - Set Flag FI1 once FC unit values have been recorded for each of P\_Hdr, P\_Data, NP\_Hdr, NP\_Data, Cpl\_Hdr, and Cpl\_Data of VLx
  - Exit to MRFC\_INIT2\_VL if:
    - ◆ Flag FI1 has been set indicating that FC unit values have been recorded for each of P\_Hdr, P\_Data, NP\_Hdr, NP\_Data, Cpl\_Hdr, and Cpl\_Data of VLx
- Rules for state MRFC\_INIT2\_VL:
- While in MRFC\_INIT2\_VL:
    - ◆ Transaction Layer must block transmission of TLPs using VLx
    - ◆ Transmit the following six MRInitFC2 DLLPs for VLx in the following relative order:
      - MRInitFC2\_VL – P – Header (first)
      - MRInitFC2\_VL – P – Data (second)
      - MRInitFC2\_VL – NP – Header(third)
      - MRInitFC2\_VL – NP – Data (fourth)
      - MRInitFC2\_VL – Cpl – Header (fifth)
      - MRInitFC2\_VL – Cpl – Data (sixth)
    - ◆ The six MRInitFC2\_VL DLLPs must be transmitted at least once every 34  $\mu$ s.
      - Time spent in the Recovery LTSSM state does not contribute to this limit.
      - It is strongly encouraged that the MRInitFC2\_VL DLLP transmissions are repeated frequently, particularly when there are no other TLPs or DLLPs available for transmission.

- ◆ Except as needed to ensure at least the required frequency of MRInitFC2 DLLP transmission, the Data Link Layer must not block other transmissions
  - Note that this includes all Physical Layer initiated transmissions (for example, Ordered Sets), Ack and Nak DLLPs (when applicable), and TLPs using VLs and VHS that have previously completed initialization (when applicable)
- ◆ Process received MRInitFC1\_VL and MRInitFC2\_VL DLLPs for VLx:
  - Ignore the indicated FC unit values
  - Set flag FI2 on receipt of any MRInitFC2\_VL DLLP for VLx
- ◆ Set flag FI2 on receipt of any MRInitFC1\_VH DLLP for any VH on VLx
- Exit to MRFC\_INIT1\_VH if:
  - ◆ Flag FI2 has been set indicating that FC unit values have been recorded for each of P\_Hdr, P\_Data, NP\_Hdr, NP\_Data, Cpl\_Hdr, and Cpl\_Data of VLx
- Rules for state MRFC\_INIT1\_VH:
  - Entered when initialization of a VH/VL (VHx VLy) is required
    - ◆ After completing MRFC\_INIT2\_VL (VHx VLy = VH0 VL0-7)
    - ◆ When some VC from a VH is mapped this VL and the indicated VC is enabled by software in the VH resulting a VH that have not been negotiated (VHx VLy = VH1-255 VL1-7, i.e. any non-zero VH on any VL)
  - If the link partner indicated no support for per-VH, VL flow control during DL\_NegotiateMR (i.e., the VH FC bit was cleared in sent MRInit DLLPs), then infinite VH credits must be advertised for all VH Credit Types.
  - While in MRFC\_INIT1\_VH:
    - ◆ Transaction Layer must block transmission of TLPs using VHx VLy
    - ◆ Transmit the following six MRInitFC1 DLLPs for VHx VLy in the following relative order:
      - MRInitFC1\_VH – P – Header (first)
      - MRInitFC1\_VH – P – Data (second)
      - MRInitFC1\_VH – NP – Header (third)
      - MRInitFC1\_VH – NP – Data (fourth)
      - MRInitFC1\_VH – Cpl – Header (fifth)
      - MRInitFC1\_VH – Cpl – Data (sixth)
    - ◆ The six MRInitFC1\_VH DLLPs must be transmitted at least once every 34  $\mu$ s.
      - Time spent in the Recovery LTSSM state does not contribute to this limit.
      - It is strongly encouraged that the MRInitFC1 DLLP transmissions are repeated frequently, particularly when there are no other TLPs or DLLPs available for transmission.

- ◆ Except as needed to ensure at least the required frequency of MRInitFC1 DLLP transmission, the Data Link Layer must not block other transmissions
  - Note that this includes all Physical Layer initiated transmissions (for example, Ordered Sets), Ack and Nak DLLPs (when applicable), and TLPs using VLs and VHS that have previously completed initialization (when applicable)
- ◆ Process received MRInitFC1\_VH and MRInitFC2\_VH DLLPs for VHx VLy:
  - Record the indicated FC unit values
  - Set Flag FI3 once FC unit values have been recorded for each of P\_Hdr, P\_Data, NP\_Hdr, NP\_Data, Cpl\_Hdr, and Cpl\_Data of VHx VLy
- Exit to MRFC\_INIT2\_VH if:
  - ◆ Flag FI3 has been set indicating that FC unit values have been recorded for each of P\_Hdr, P\_Data, NP\_Hdr, NP\_Data, Cpl\_Hdr, and Cpl\_Data of VHx VLy
- Rules for state MRFC\_INIT2\_VH:
  - While in MRFC\_INIT2\_VH:
    - ◆ Transaction Layer must block transmission of TLPs using VHx VLy
    - ◆ If the link partner indicated no support for per-VH, VL flow control during DL\_NegotiateMR (i.e., the VH VL bit was cleared in sent MRInit DLLPs), then infinite VH credits must be advertised for all VH Credit Types.
    - ◆ Transmit the following six MRInitFC2 DLLPs for VHx VLy in the following relative order:
      - MRInitFC2\_VH – P – Header (first)
      - MRInitFC2\_VH – P – Data (second)
      - MRInitFC2\_VH – NP – Header (third)
      - MRInitFC2\_VH – NP – Data (fourth)
      - MRInitFC2\_VH – Cpl – Header (fifth)
      - MRInitFC2\_VH – Cpl – Data (sixth)
    - ◆ The six MRInitFC2\_VH DLLPs must be transmitted at least once every 34  $\mu$ s.
      - Time spent in the Recovery LTSSM state does not contribute to this limit.
      - It is strongly encouraged that the MRInitFC2\_VH DLLP transmissions are repeated frequently, particularly when there are no other TLPs or DLLPs available for transmission.
    - ◆ Except as needed to ensure at least the required frequency of MRInitFC2 DLLP transmission, the Data Link Layer must not block other transmissions
      - Note that this includes all Physical Layer initiated transmissions (for example, Ordered Sets), Ack and Nak DLLPs (when applicable), and TLPs using VLs and VHS that have previously completed initialization (when applicable)
    - ◆ Process received MRInitFC1\_VH and MRInitFC2\_VH DLLPs for VHx VLy:

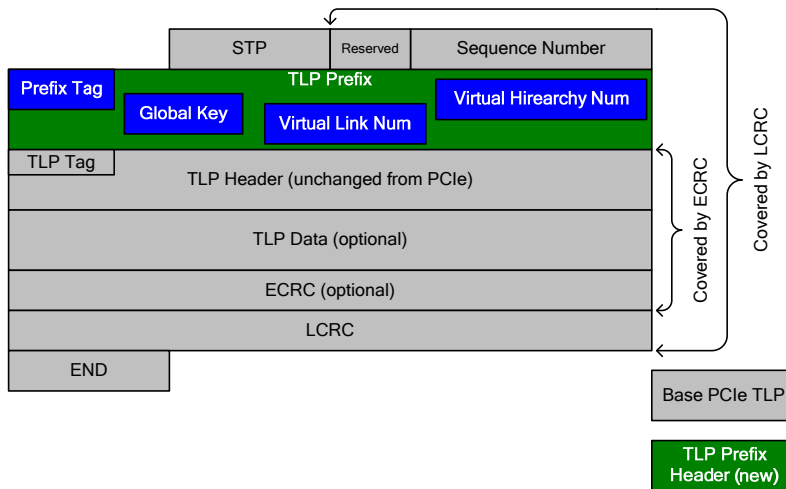


- Ignore the indicated FC unit values
- Set flag FI4 on receipt of any MRInitFC2\_VH DLLP for VHx VLy
- ◆ Set flag FI4 on receipt of any TLP on VHx VLy, or any UpdateFC DLLP for VHx VLy
- ◆ Signal completion and exit if:
  - Flag FI4 has been set

## 2.2. TLP Prefix Tagging

After successful MR Link Protocol negotiation, links use the MR Enhanced Link Protocol. TLPs on such links contain a TLP Prefix as shown in [Figure 2-16](#). This prefix is located between the Sequence Number and the PCIe TLP Header.

Deleted: Figure 2-16



**Figure 2-16: TLP Prefix Header Location**

The TLP Prefix is part of the wire packet. It is covered by LCRC and is retransmitted with the rest of the TLP under the same rules. Sequence Numbers and the Ack/Nak protocol remain a per-link notion and are not affected by Multi-Root operation.

The TLP Header, Data and ECRC are not changed from PCIe TLP usage.

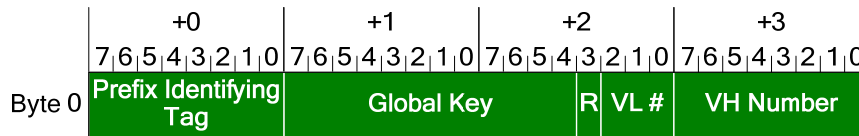


Figure 2-17: TLP Prefix Header Layout

The layout of the TLP Prefix Header is shown in [Figure 2-17](#). The first byte is a fixed Prefix Identifying Tag value. This value is not used by any TLP (now or in the future) so that the presence of the TLP Prefix can be determined by examining the TLP in isolation without needing additional link-state information. Doing so allows test equipment to better understand things without needing to observe the initialization sequence.

Virtual Hierarchy Numbers (VHN) are link-local. On a link that supports  $n$  VHs, valid VHNs are  $[0 .. n-1]$ . A given TLP may take on different VHN values on each Multi-Root link that it traverses.

The VL # field contains the Virtual Link Number (VL #). It contains information used to support Congestion Management and Isolation. TLPs are assigned to VLs using a combination of PCIe TC to VC mapping rules and new MR VC to VL mapping rules. See Chapter 8 Congestion Management for details.

The Global Key field is used to guard against “VH Hopping”. VH Hopping is when a TLP in one VH inadvertently ends up on a different VH (either due to MR-PCIM table configuration errors or due to a hardware error inside a MR Switch). The Global Key is added to the TLP when the TLP Prefix is attached at the MR Ingress point. The Global Key value selected is based on which VH is being used. This Global Key value is preserved, unchanged, through subsequent MR Switches. The Global Key value is validated against the expected value at various points in the MR topology. The Global Key is always validated at MR Egress. The Global Key value is optionally validated at either (or both) MR Switch Input or MR Switch Output. Global Key checking is similar to PCIe ECRC checking. Failure of any Global Key validation is an unrecoverable error.

MR-PCIM software configures the tables used to generate and check the Global Key values. To avoid Global Key mismatch errors, MR-PCIM must configure tables such that all TLPs in a given VH have the same Global Key value. To provide maximum protection, MR-PCIM should configure tables so that TLPs in different VHs have different Global Keys (this protection is not provided between VHs with duplicate Global Key values).

## 2.2.1. MR Switch Transaction Layer Processing

MR Switches implement a set of Virtual Switches (VS). Each VS is assigned by software to a single Virtual Hierarchy that, in turn, is associated with a single Root Port. Within a Switch, TLPs are associated with a VS and are routed within that VS using PCIe routing and ordering rules (e.g. address routed, ID routed, broadcast to/from root). TLPs for unrelated VHs are unordered.

TLPs are processed by a switch as follows:

1. TLPs arrive on an Input Port of the switch with a link-local Input VH Number.
  - a. For a link operating in MR mode, the Input VHN is contained in the TLP Prefix Header.

Deleted: Figure 2-17

**Comment [sdg1]:** Coordination with the Protocol Working Group will occur to define the value and appropriate steps will be taken to ensure that the selected value remains “unused” forever.

**Comment [m2]:** A VP properties list should be used so that all attributes are maintained in one place. This may be part of chapter 1.

- b. For a link operating in Base PCIe mode the Input VHN is 0.
2. Switch mapping tables are used to map the incoming TLP to a VS and a specific PCI-to-PCI Bridge of that VS.
 
$$f(\text{Input Port, Input VHN}) \rightarrow \{\text{VS, Input Bridge}\}$$
3. Global Key input processing occurs.
  - a. If the input link was operating in MR mode, the Global Key from the TLP is optionally validated against the global key associated with the VS. This is the “entering check” as described in Section 4.3.5.2.
  - b. If the input link was operating in Base PCIe mode, there was no Global Key to check against. This is the MR Ingres point for this TLP and the TLP is assigned the Global Key associated with the VS.
4. The TLP is routed within the VS. This mapping uses conventional PCIe rules using mapping tables controlled by software operating in the VH. If the TLP is consumed by the VS, these rules indicate routing within the VS (i.e. which Type 1 Header, etc.). If the TLP is being forwarded onward, these rules select an Output Bridge and an output Virtual Circuit (VC).
 
$$\begin{aligned} g(\text{Input Bridge, Base TLP Header}) &\rightarrow \{\text{Type 1 Header ...}\} && \text{if TLP is local to the VS} \\ &\rightarrow \{\text{Output Bridge, VC}\} && \text{if TLP is being forwarded} \end{aligned}$$
5. For TLPs being consumed by the VS or being forwarded to an Output Port operating in Base PCIe mode, this switch is the MR Egress point for the TLP and Global Key checking occurs. The Global Key associated with the TLP is optionally validated against the Global Key associated with the VS. This is the “terminating check” as described in Section 4.3.5.2.
6. For TLPs being forwarded, the switch mapping tables are used to map the outgoing TLP to an Output Port, an Output VHN of that Port and an Output Virtual Link (VL) of that Port
 
$$h(\text{VS, Output Bridge, VC}) \rightarrow \{\text{Output Port, Output VHN, Output VL}\}$$
7. The VL and (VH, VL) Flow Control gates are checked using the Output Port’s values to verify that there are sufficient credits to forward the TLP.
8. Port Arbitration is performed using PCIe rules within the VS. Each Downstream Bridge is considered a distinct Port for this purpose.
9. For an output link operating in MR mode, VH Arbitration occurs within the Output Port, Output VL. Arbitration scheme used is fixed round robin.
10. For an output link operating in MR mode, VL Arbitration occurs within the Output Port. Arbitration scheme is controlled by the VL Arbitration information in the Port Table.
11. For an output link operating in Base PCIe mode, VC Arbitration occurs within the Output Port. Arbitration scheme is controlled by software using VC Arbitration information associated with the Type 1 Header within the VS. Programmable VC Arbitration is optional in PCIe and remains optional for links operating in Base PCIe mode. VC Arbitration is not supported for links operating in MR mode.
12. The Output Port forwards the TLP.

- a. For an output link operating in Base PCIe mode, the Output VHN is zero. The Virtual Circuit (VC) used to transmit the TLP was determined in step 4 above. The Output VL value is not used.
  - b. For an output link operating in MR mode new Output VHN and Output VL values are placed in the TLP Prefix Header overwriting the input values (if any). The Output VL is used to transmit the TLP.
13. Global Key output processing occurs. If the output link is operating in MR mode, the Global Key from the TLP is optionally validated against the global key associated with the VF. This is the “exiting check” as described in Section 4.3.5.2. If the output link is operating in Base PCIe mode, the “terminating check” in step 5 is used instead.
  14. When receive buffer space is made available, Flow Control is returned to the VL and (VH, VL) contained in the TLP Prefix. The TC to VC maps and VC to VL maps on the receiver are not used (PCIe does not transmit the VC so the TC to VC map is used for this purpose).

## 2.2.2. MR Device Transaction Layer Processing

MR Devices implement a collection of Functions in each VH. VH0 is used to manage the device. Within a Device, TLPs are associated with a VH and are routed within that VH using PCIe routing and ordering rules. TLPs for unrelated VHS are unordered.

**Comment [m3]:** A VP properties list should be used so that all attributes are maintained in one place. This may be part of chapter 1.

### 2.2.2.1. Receiving TLPs

TLPs received by a MR Device are processed as follows:

1. TLPs arrive from the MR Switch with a link-local Input VH Number contained in the TLP Prefix Header.
2. The Global Key from the TLP is validated against the global key associated with the VH.
3. The addressed Function is determined using PCIe rules *applied within the VH*.
  - a. If the TLP is ID routed, the captured Bus # and the TLP's ID are used.
  - b. If the TLP is Address routed, these various BARs and the TLP's address are used.
4. The addressed Function is one of PF, VF, Function or BF. These are designated as **PF h:f**, **VF h:f,s**, **F h:f** or **BF bf** respectively (using the nomenclature described in Section 1.2).
5. PCIe TC checking occurs. For VFs, the PF capabilities are used.
  - a. If the Function has a VC Capability, the TC of the TLP is checked to verify that the TC is enabled in the VC Capability.
  - b. If the Function does not have a VC Capability, the TC must be 0.
  - c. If Function 0 in the VH supports has a MFVC, the TC of the TLP is checked to verify that the TC is enabled in the MFVC Capability.

6. The VH number *h* and PF/VF/Function number *f* determine the associated BF and corresponding Function Table Entry of that BF. For unmanaged Functions in VH0, there will not be an entry.
7. For Functions, BFs and PFs, a Vendor Specific mechanism determines the underlying Function.
8. For VFs, VF Mapping is used to determine the underlying Function
  - a. If the BF located in step 6 supports VF Mapping, the MVF number is contained in the LVF Table Entry located using the VF # together with the TotalVFs and BaseLVF values from the Function Table Entry. This MVF number describes the underlying Function.
  - b. In addition if VF Migration is supported and enabled, the VF State in the LVF Table Entry must also be checked to ensure the mapping is in either of the Active.Available or Active.MigrateOut states.
  - c. If the BF located in step 6 does not support VF Mapping, a Vendor Specific Mechanism determines the underlying Function.
9. The TLP is handed to underlying Function for processing.
10. When receive buffer space is made available, Flow Control is returned to the VL and (VH, VL) contained in the TLP Prefix. The TC to VC maps and VC to VL maps on the receiver are not used (PCIe does not transmit the VC so the TC to VC map is used for this purpose).

#### 2.2.2.2. *Transmitting TLPs*

TLPs transmitted by a MR Device are processed as follows:

1. The underlying Function, TC, TLP size and TLP type (P, NP, Cpl) are determined.
2. For Functions, BFs and PFs a Vendor Specific mechanism determines the VH and Function #.
3. For VFs, if VF Mapping is used, a reverse VF map is used to determine the VH and Function #. Otherwise, a Vendor Specific mechanism determines the VH and Function #.
4. Initial TC to VCID mapping and VC Arbitration occurs. The Mapped Function's VC Capability is used to convert TC to VCID. For VFs, the VC Capability of the PF is used.
5. If Function 0 of the VH supports the MFVC Capability, MFVC Arbitration and another round of TC to VCID mapping occurs.
6. VCID to VL mapping occurs using the map contained in the Function Table Entry associated with the Function or PF.
7. Now that the VL is known, VL and (VH, VL) Flow Control gates checking occurs.
8. VH Arbitration occurs with the VL. The arbitration scheme is fixed round-robin.
9. VL Arbitration occurs. The VL Arbitration tables are contained in one of the BFs.
10. The TLP is sent on the wire. CREDITS\_CONSUMED is updated to reflect this.

### 2.2.3. Global Key Processing

Global Keys are processed at various points within an MR topology:

- ❑ Global Keys are assigned at the MR Ingress point of the TLP. For TLPs originated by an MR Component, the MR Ingress point is inside that originating MR Component. For TLPs originated by a Base PCIe Component, the MR Ingress point is the input port of an MR Switch.
- ❑ Global Keys are checked at the MR Egress point of the TLP. For TLPs destined for an MR Component, the MR Egress point is inside the destination MR Component. For TLPs destined for a Base PCIe Component, the MR Egress point is the output port of an MR Switch.
- ❑ MR Switches can optionally check Global Keys as TLPs enter the switch from an MR link.
- ❑ MR Switches can optionally check Global Keys as TLPs exit the switch on an MR link.

Hardware support for the check as TLPs enter and exit MR Switches is optional. Hardware support for the MR Egress check is mandatory.

The Global key value is a 12 bit value, assigned by software to each VH. To achieve maximum protection, software should assign each VH a distinct Global Key value.

A Global Key check passes if the value in the TLP and the expected value match. A Global Key check also passes if either the expected value is 000h or the TLP value is 000h (the wild card value).

In other words, when MR Ingress points assign a TLP the Global Key value of 000h, Global Key checks will always pass for that TLP. When a Global Key is programmed with an expected value of 000h, any checks that use that value will always pass.

Global Key checking is disabled by default. This allows software time to program the Global Key registers before enabling checking.

## 2.2.4. MR TLP Dataflow Examples

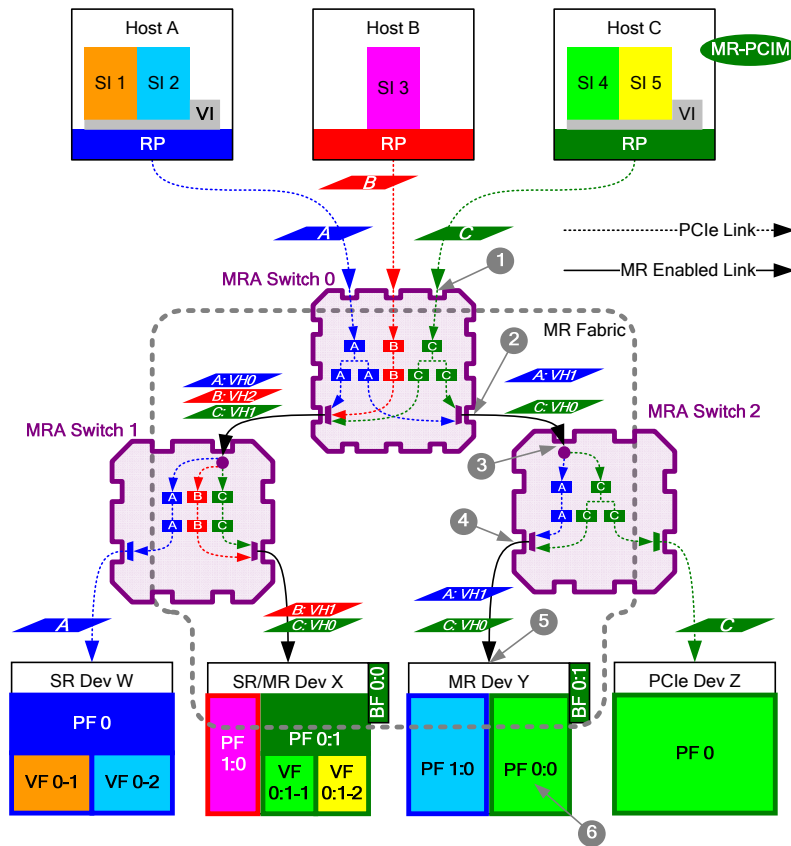


Figure 2-18: MR Dataflow Examples

Consider a Memory Read TLP initiated by SI 4 targeting PF 1:0 in EP Y:

1. The Host C to MRA Switch 0 link operated in PCIe mode. There is no TLP Prefix.
2. MRA Switch 0 has been programmed so that all TLPs arriving at Port ❶ are associated with the VS for VH C.
3. The VS inside MRA Switch 0 address routes the TLP to a Virtual Downstream Port associated with physical Port ❷ (the link headed towards MRA Switch 2). MR-PCIM has assigned VH 0 to this Virtual Downstream Port. The TLP exits MRA Switch 0 with a TLP Prefix having VH Number 0.

4. The TLP arrives at MRA Switch 2 Port ③ labeled as belonging to VH 0. MRA Switch 2 has been programmed so that all TLPs arriving at Port ③ labeled VH 0 are associated with the VS for VH C.
5. The VS inside MRA Switch 2 address routes the TLP to a Virtual Downstream Port associated with physical Port ④ (the link headed towards Device Y). MR-PCIM has assigned VH 1 to this Virtual Downstream Port. The TLP exits MRA Switch 2 with a TLP Prefix having VH Number 1.
6. The TLP arrives at Device Y labeled as belonging to VH 1. The Device hands the transaction to PF 1:0 ⑥ for execution.
7. PF 1:0 completes the transaction and emits a completion TLP. Device Y sends this TLP out Port ⑤ labeled with VH 1.
8. The TLP arrives at MRA Switch 2 Port ④ labeled as belonging to VH 1. MRA Switch 2 has been programmed so that all TLPs arriving at Port ④ labeled VH 1 are associated with the VS for VH C.
9. The VS inside MRA Switch 2 ID routes the TLP to a Virtual Upstream Port associated with physical Port ③ (the link headed towards MRA Switch 0). MR-PCIM has assigned VH 0 to this Virtual Upstream Port. The TLP exits MRA Switch 2 with a TLP Prefix having VH Number 0.
10. The TLP arrives at MRA Switch 0 Port ② labeled as belonging to VH 0. MRA Switch 0 has been programmed so that all TLPs arriving at Port ② labeled VH 0 are associated with the VS for VH C.
11. The VS inside MRA Switch 0 ID routes the TLP to a Virtual Upstream Port associated with physical Port ① (the link headed towards Host C). MR-PCIM has designated this link a PCIe link and has assigned this Virtual Upstream Port to it. The TLP exits Switch 0 with a no TLP Prefix.
12. SI 4 in Host C sees a completion for the Memory Read Transaction.

## 2.3. Per-VH RESET

In Multi-Root, Reset DLLPs are used to implement a per-VH equivalent of PCIe Hot Reset. DLLPs are used to ensure “quick” Reset propagation by avoiding delays introduced by TLP ordering and flow control rules.

Reset of a VH requires discarding TLPs associated with that VH. TLPs associated with VHs not in reset must not be affected. This discard process may take more time than the PCIe equivalent. For example, only some of TLPs in the retry buffer might be affected.

Reset DLLPs supports an acknowledgment protocol to mimic the TS1/TS2 ACK behavior of PCIe. This allows an upstream component to know when a downstream component has entered reset.



### 2.3.1. Per-VH Reset Example

Figure 2-19 shows an example MR Topology. Figure 2-20 shows Host A's view of the same topology. Table 2-3 describes the events and associated actions involved in propagating a Reset from Host A.

Deleted: Figure 2-19

Deleted: Figure 2-20

Deleted: Table 2-3

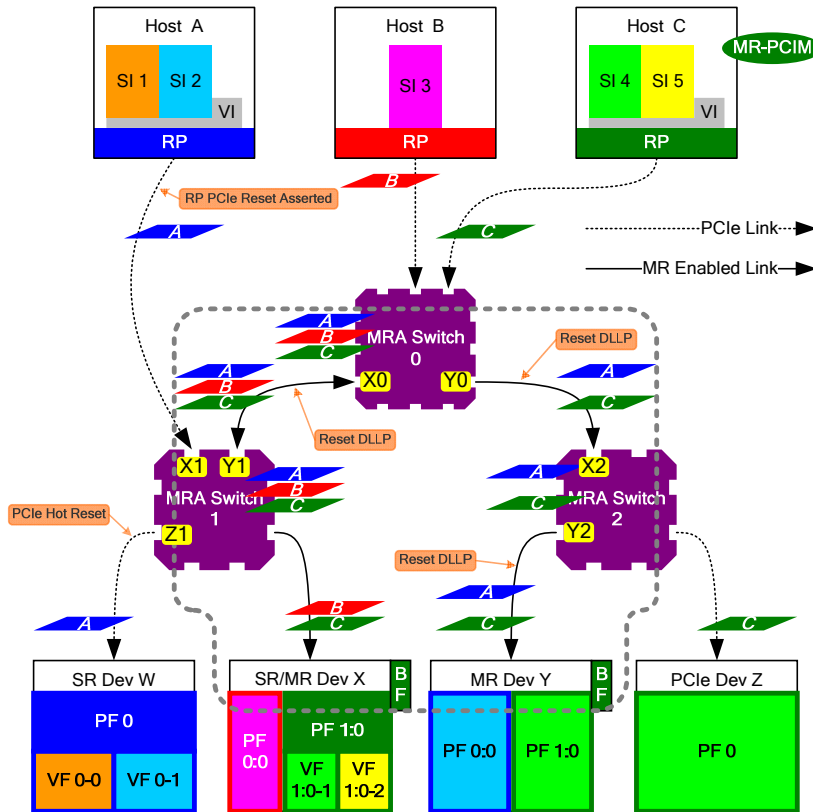
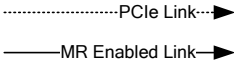


Figure 2-19: Reset DLLP Example: Topology



**Figure 2-20: Reset DLLP Example: Host A's View**

**Table 2-3: Reset DLLP Example: Events and Actions**

Component	Event	Action
Host A	Sends Hot Reset on $RP_A$ to MRA Switch 1	
MRA Switch 1	Sees Hot Reset on Port X1 which is the Upstream Port of VH A	1.1 Discards all TLPs headed out Port X1 1.2 Acks Hot Reset on Port X1 (PCIe TS1/2) 1.3 Sends Hot Reset out Port Z1 to EP W 1.4 Sends Reset DLLP Request for VH A out Port Y1 to MRA Switch 0 <i>Unlabeled Ports not affected</i>
	All TLPs for VH A are discarded or marked for discard and Port X1 Retry Buffer empty	1.5 Nothing – VH A Upstream Port is not MR Link
	Sees Reset Ack DLLP on VH A from Port Y1	1.6 Knows MRA Switch 0 will not send any more TLPs for VH A through Port Y1 1.7 Stops resending Reset DLLP for VH A out Port Y1
MRA Switch 0	Sees Reset DLLP Request for VH A on Port X0 which is the Upstream Port of VH A	0.1 Starts discarding TLPs for VH A 0.2 Sends Reset DLLP Request for VH A out Port Y0 to MRA Switch 2 <i>Unlabeled Ports not affected</i>
	All TLPs for VH A are discarded or marked for discard and Port X0 Retry Buffer contains no TLPs for VH A	0.3 Send Reset Ack DLLP out Port X0 to MRA Switch 1
	Sees Reset Ack DLLP on VH A from Port Y0	0.4 Knows MRA Switch 2 will not send any more TLPs for VH A through Port Y0 0.5 Stops resending Reset DLLP for VH A out Port Y0
MRA Switch 2	Sees Reset DLLP Request for VH A on Port Z0 which is the Upstream Port of VH A	2.1 Starts discarding TLPs for VH A 2.2 Sends Reset DLLP Request for VH A out Port Y2 to EP Y <i>Unlabeled Ports not affected</i>
	All TLPs for VH A are discarded or marked for discard and Port X2 Retry Buffer contains no TLPs for VH A	2.3 Send Reset Ack DLLP out Port X0 to MRA Switch 1
	Sees Reset Ack DLLP on VH A from Port Y2	2.4 Knows EP Y will not send any more TLPs for VH A through Port Y2 2.5 Stops resending Reset DLLP for VH

		A out Port Y2
Device W	Sees PCIe Hot Reset	W.1 Discards all TLPs, enters Reset W.2 Acks Hot Reset (PCIe TS1/2)
Device Y	Sees Reset DLLP Request on VH A	Y.1 Starts discarding TLPs for VH A
	All TLPs for VH A are discarded or marked for discard and no TLPs for VH A are in the Retry Buffer	Y.2 Send Reset Ack DLLP
Device X and Device Z	Nothing, not part of VH A	

### 2.3.2. RESET DLLP Format

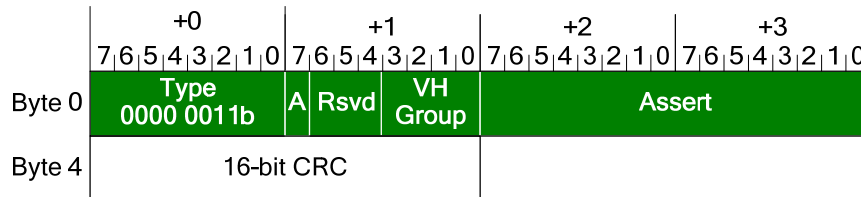
Figure 2-21 shows the bit encoding of the Reset DLLP.

Deleted: Figure 2-21

The A bit is 0 for a Reset Request (propagating downstream) and is 1 for a Reset Ack (propagating upstream).

The VH Group contains the upper bits of the VHN.

The Assert field contains one bit for each of 16 VHs within a VH Group. An Assert bit is 1 to indicate that the associated VH is in Reset, and a 0 to indicate that the associated VH is not in Reset.



**Figure 2-21: Reset DLLP**

Reset DLLPs can be sent at any time a link is in DL\_Active. Reset Request DLLPs request a downstream link partner enter or exit Reset state on one or more VHs. Reset Ack DLLPs indicate that the downstream link partner has seen and processed the reset request.

Sending a Reset DLLP with the assert bit set is making a promise about flushing of stale TLPs. In particular, a component sending a Reset Request or Ack DLLP with Assert == 1 is claiming that all TLPs from before the VH entered Reset have been flushed (either discarded or marked for later discard). This claim includes TLPs that are sitting in the Retry buffer. Consequently, typical implementations will delay sending the Reset DLLP until the affected TLPs are retired from the Retry Buffer using normal Ack protocol (TLPs transmitted during this time will be discarded at the remote end of the link).

### 2.3.3. RESET DLLP Processing

The following sections describe Reset DLLP processing steps in an upstream and downstream ends of a link. Recall that upstream and downstream are relative to a VH. In particular, one end of a link

may be using the upstream state machine for some VHS and the downstream state machine for other VHS.

The upstream and downstream state machines run in parallel on every VH.

### 2.3.3.1. Upstream State Machine

The following steps are used to enter Reset. This is triggered by a request to send a Reset DLLP on a link. This request can occur for a variety of reasons (e.g. DL\_DOWN, Reset DLLP Request, or Hot Reset on an upstream Switch link, setting Secondary Bus Reset bit in some Type 1 configuration header, etc.)

1. Reset Requested.
2. Start discarding new TLPs received for this VH from this link.
3. Start discarding new TLPs to be transmitted for this VH on this link.
4. Discard or mark for discard any TLPs waiting to be sent for this VH on this link.
5. Wait until the Retry Buffer contains no TLPs for this VH (i.e. they have been acknowledged and thus removed from the Retry Buffer).<sup>1</sup>
6. Schedule a Reset Request DLLP to be sent with the Assert bit = 1
7. Schedule resending the Reset Request DLLP approximately every 30  $\mu$ sec until a Reset Ack DLLP with the Assert bit = 1 is received (see Section 2.3.3.3 for details).
8. If a Reset Ack DLLP is received with Assert bit = 1, the remote end has entered Reset, stop scheduling Reset Request DLLPs for this VH (Reset Request DLLPs may continue to be sent if needed by other VHS).

The following steps are used to exit Reset. This is triggered by the condition causing the entry into Reset going away (e.g. clearing Secondary Bus Reset, Physical LinkUp transitions from 0 to 1 etc.).

1. Schedule a Reset Request DLLP to be sent with the Assert bit = 0
2. Schedule resending the Reset Request DLLP approximately every 30  $\mu$ sec until Reset Ack DLLP with the Assert bit = 0 is received (see Section 2.3.3.3 for details).
3. If a Reset Ack DLLP has been received with Assert bit = 0, the remote end has exited Reset, stop scheduling Reset Request DLLPs for this VH (Reset Request DLLPs may continue to be sent if needed by other VHS).

Reset Request DLLPs may be coalesced so that multiple schedule events result in a single DLLP being transmitted.

Timely propagation of Reset is important. Components should send Reset Request DLLPs as soon as possible after detecting after starting to enter or exit Reset.

---

<sup>1</sup> This could be implemented by waiting until all TLPs, for any VH, have been flushed from the Retry Buffer that were in the Retry Buffer prior to Step 3 (Step 3 ensures that no new TLPs for this VH will enter the Retry Buffer).

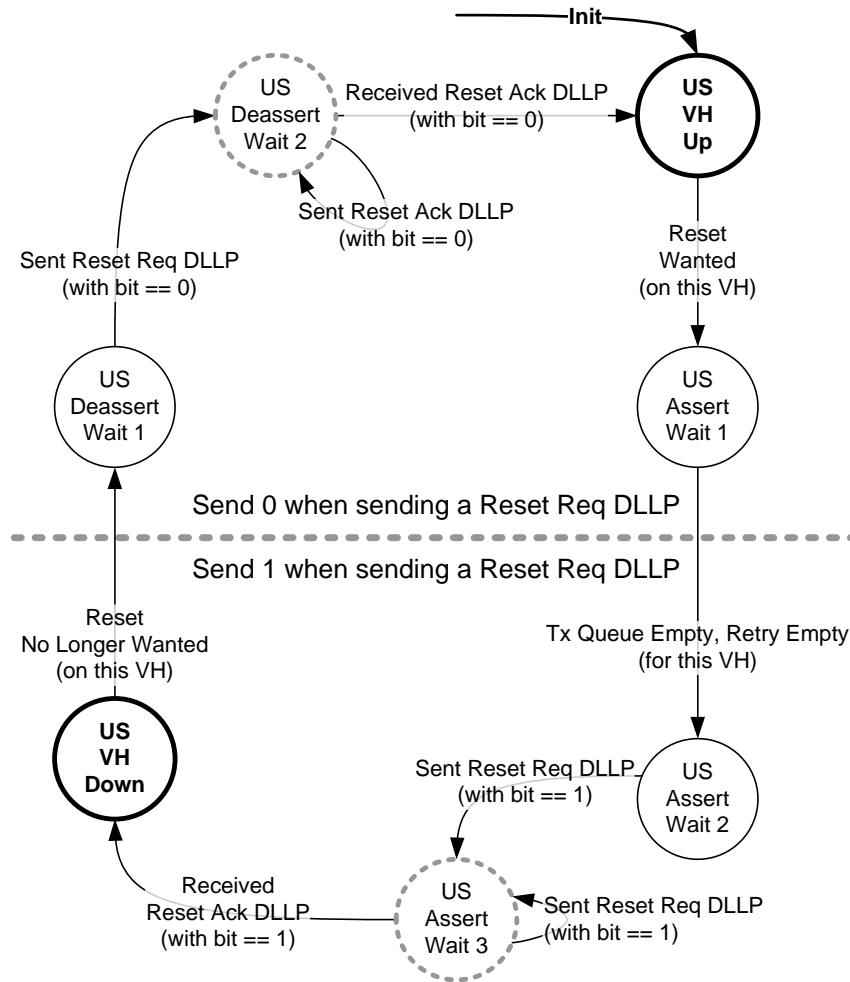


Figure 2-22: Upstream Link Partner RESET SM

### 2.3.3.2. Downstream State Machine

The following steps are used to enter Reset. This is triggered receiving a Reset DLLP on a link with the Assert bit = 1.

1. Reset Request DLLP is received with the Assert bit = 1
2. Start discarding new TLPs received for this VH from this link.
3. Start discarding new TLPs to be transmitted for this VH on this link.

4. Discard or mark for discard any TLPs waiting to be sent for this VH on this link.
5. All Functions in the VH enter the Reset
6. Wait until the Retry Buffer contains no TLPs for this VH (i.e. they have been acknowledged and thus removed from the Retry Buffer).
7. Schedule a Reset Ack DLLP to be sent with the Assert bit = 1
8. Schedule another Reset Ack DLLP to be sent whenever a Reset Request DLLP is received and the Assert bits in the Reset Request match what would be transmitted in the Reset Ack (i.e. retransmit the Ack in case it got lost).

The following steps are used to exit Reset. This is triggered by receiving a Reset DLLP on a link with the Assert bit = 0.

1. Reset Request DLLP is received with the Assert bit = 0
2. If all Functions are ready to exit Reset, Schedule a Reset Ack DLLP to be sent with the Assert bit = 0
3. Schedule another Reset Ack DLLP to be sent whenever a Reset Request DLLP is received.

Reset Ack DLLPs may be coalesced so that multiple schedule events result in one DLLP being transmitted.

Timely acknowledgement of Reset is important. Components shall respond with Reset Ack within 1.9 ms (+0%, -100%) and are strongly encouraged to respond much quicker. The 1.9 ms value is chosen to avoid inadvertent link retraining caused by the Reset DLLP Forward Progress Timer (see section 2.3.3.3).

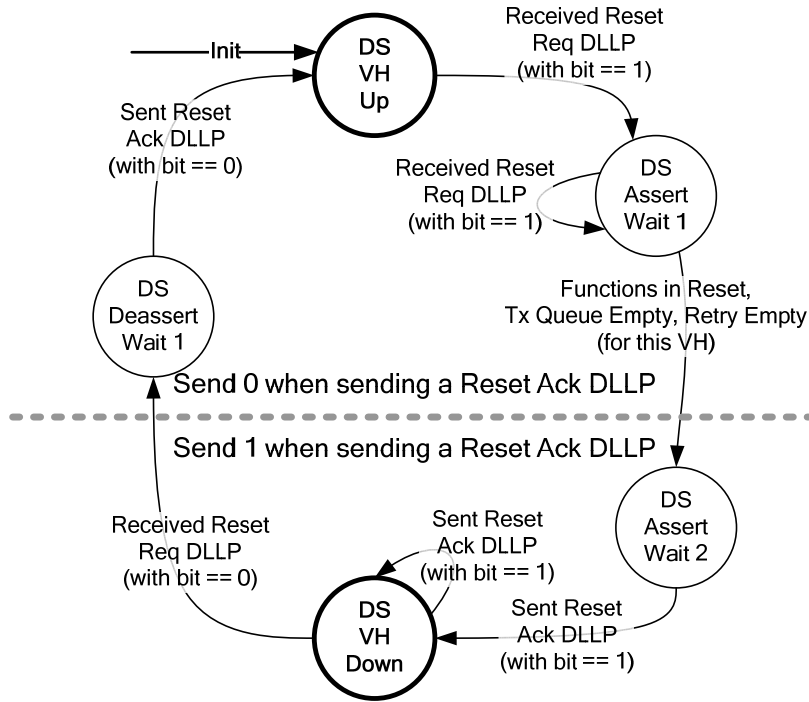


Figure 2-23: Downstream Link Partner RESET SM

### 2.3.3.3. Reset DLLP Reliability

A VH is considered to be waiting for Reset Ack when it is states US Assert Wait 3 or US Deassert Wait 2.

The Upstream component retransmits Reset Requests approximately every 30  $\mu$ sec as controlled by the Reset Request Retransmit timer. This timer is enabled whenever any VH is waiting for Reset Ack. It restarts when a Reset Request DLLP is transmitted for any VH group (either initial DLLP or a resend). It expires 30  $\mu$ sec after being started (+50%, -0%). When the timer expires, Reset Request DLLPs are scheduled to be sent for all VH groups that have some VH waiting for a Reset Ack.

The Upstream component also includes a Reset DLLP Forward Progress timer. This timer is enabled when any VH is waiting for Reset Ack. It restarts when some VH enters either waiting for Reset Ack state or a Reset Ack is received for any VH. It expires 2 ms after being started (+50%, -0%). When the timer expires, a link retrain is requested.

The Upstream component also includes a 2 bit Reset Retrain counter. This counter is incremented when a link retrain is requested due to the expiration of the Reset DLLP Forward Progress timer. This counter resets to 00b whenever the Reset DLLP Forward Progress timer is restarted. If this counter rolls over from 11b to 00b, the link shall enter enters Detect.



Reset state machines are not affected by link retraining (either initiated by the Reset DLLP Forward Progress Timer or through other means). DLLPs that were scheduled during the link retrain shall be sent when the link retrain completes.

Requests to retrain the link initiated by this mechanism may be coalesced with requests to retrain the link initiated by other mechanisms. For example, two “simultaneous” requests for a link retrain due to (1) a REPLAY\_NUM rollover and to (2) the Reset Forward Progress timer may result in either one or two retrain sequences.

#### *2.3.3.4. Flow Control and Reset / DL\_DOWN*

In Multi-Root systems, Flow Control DLLPs can affect more than one VH. It is critical that a VH entering or exiting the Reset state does not disrupt other VHs. Flow Control credits on MR Enabled Links must be returned to the originator even when the VH that originated the discarded TLPs is in or is entering Reset.

For example, suppose TLPs enter a Switch on the MR Link associated with Port 1 destined for Port 2. If Port 2 enters Reset or DL\_DOWN and thus discards these TLPs, credits must be returned in the appropriate VH of Port 1. This must occur even if the associated VH at Port 1 is also in or is entering reset (there may be other VHs not in Reset sharing Port 1). This must occur whether or not Port 2 is an MR Link.

Reset propagation may not be affected by Flow Control. Reset DLLPs must be sent independent of any Flow Control state.

Reset propagation is affected by the TLP Ack / Nak protocol. Allowing this avoids the complexity of editing the Retry Buffer to remove TLPs for VHs that are now in Reset.

## 2.4. MR Flow Control

### 2.4.1. FC Information Tracked by Transmitter

A transmitter shall track the following two quantities for each supported (VH, VL) and VL. As in PCIe, Header and Data credits are tracked independently.

- ❑ CREDITS\_CONSUMED
  - Computed and used in the same manner as in PCIe protocol.
- ❑ CREDIT\_LIMIT
  - Use in the same manner as in PCIe protocol
  - Undefined at interface initialization.
  - Set to the value indicated during MR Flow Control initialization
  - Updated based on MRUpdateFC DLLPs as described below.

In MR, transmitting a TLP requires passing 4 gates instead of the 2 gates used in PCIe.

- ❑ One PH, NPH or CPLH header VL credit is required.
- ❑ One PH, NPH or CPLH header (VH, VL) credit is required.
- ❑ PD, NPD or CPLD data VL credits may be required. The number of credits needed depends on the TLP size and uses the same rules as PCIe.
- ❑ PD, NPD or CPLD data (VH, VL) credits may be required. The number of credits needed depends on the TLP size and uses the same rules as PCIe.

The transmitter gating function for a TLP is said to pass if the transmitter gating function passes for all four gates. If any gate fails, the transmitter must block transmission of the TLP. If CREDIT\_LIMIT was specified as “infinite” during Flow Control initialization, then the corresponding gating function is unconditionally satisfied for that type of credit.

The Transmitter must following the same ordering and deadlock avoidance rules as specified in the PCIe protocol. TLPs mapped to different VLs have no ordering relationship, and must not block each other.

The transmitter gating function rules for (VH, VL)s and VLs are the same as that specified in the PCIe protocol.

TLPs associated with different VHs represent different flows and have no ordering relationship. TLPs blocked due to failure of the (VH, VL) gating function should not block TLPs associated with other VHs mapped to the same VL.

If any CREDIT\_LIMIT\_VHVL for a Credit Type is infinite, then all CREDIT\_LIMIT\_VHVL Credit Type values for that VL must also be infinite.

If the CREDIT\_LIMIT\_VL and CREDIT\_LIMIT\_VHVL are both infinite, no UpdateFC or MRUpdateFC DLLPs are sent. A transmitter may optionally raise a Receiver Error if one is received (as in PCIe).

Otherwise, if CREDIT\_LIMIT\_VHVL is non-infinite, then MRUpdateFC DLLPs are sent.

- ❑ MRUpdateFC DLLPs update the values of VL and (VH, VL) CREDIT\_LIMIT as follows:
  - ◆  $\text{Credits\_Received} = (\text{Update Value} - \text{CREDIT\_LIMIT\_VHVL}) \bmod 2^{\text{Field Size}}$
  - ◆  $\text{CREDIT\_LIMIT\_VL} = (\text{CREDIT\_LIMIT\_VL} + \text{Credits\_Received}) \bmod 2^{\text{Field\_Size}}$
  - ◆  $\text{CREDIT\_LIMIT\_VHVL} = \text{Update Value}$
  - ◆ The CREDIT\_LIMIT\_VL value is updated using the CREDIT\_LIMIT\_VHVL value *before* it is updated by this DLLP.
- ❑ The transmitter computes but otherwise ignores the CREDIT\_LIMIT values associated with infinite CREDIT\_LIMIT\_VL credits.

Otherwise, if CREDIT\_LIMIT\_VHVL is infinite, then PCIe Base UpdateFC DLLPs are sent. A transmitter may optionally raise a Receive Error if an MRUpdateFC DLLP is received.

- ❑ UpdateFC DLLPs update the value of VL CREDIT\_LIMIT as follows:
  - ◆  $\text{CREDIT\_LIMIT\_VL} = \text{Update ValueFC}$

**Comment [po4]:** PCIe still allows UpdateFCs to be sent. Need to resolve how to handle this.

#### 2.6.1. Flow Control Rules

If an Infinite Credit advertisement (value of 00h or 000h) has been made during initialization, no Flow Control updates are required following initialization.

- If UpdateFC DLLPs are sent, the credit value fields must be set to zero and must be ignored by the Receiver. The Receiver may optionally check for non-zero update values (in violation of this rule). If a component implementing this check determines a violation of this rule, the violation is a Flow Control Protocol Error (FCPE)

## 2.4.2. Information Tracked by Receiver

A receiver shall track the following two quantities for each supported (VH, VL) and VL.

### ❑ CREDITS\_ALLOCATED

- The initial value is Vendor Specific. This value was communicated to the transmitter using the MR Flow Control Initialization Protocol
- This value is incremented as processing (or discarding) of received TLPs makes additional receiver buffer space available.
- Changes to this value are communicated to the transmitter using flow control update DLLPs.

### ❑ CREDITS\_RECEIVED

- Computed in the same manner as in PCIe protocol.

It is a goal to reduce the amount of FCP traffic associated with inactive flows. After an MRUpdateFC has been sent 4 times with the same value, subsequent MR UpdateFC FCPs with the same value must be scheduled for transmission once every 100 ms (-0%/+50%).

If a Receiver advertises non-infinite VL credits, then it must send MRUpdateFC DLLPs whenever an FCP for that VL must be scheduled for transmission.

If a Receiver advertises infinite VH credits, then it must send UpdateFC DLLPs whenever an FCP for that VL must be scheduled for transmission.

## 2.5. MR Message Processing

### 2.5.1. Interrupts

Interrupt processing occurs within a VH. The associated TLPs contain a TLP Prefix allowing all components to route them appropriately.

MSI and MSI-X Interrupts are indistinguishable from other Memory Write TLPs.

INTx Interrupts are represented using ASSERT\_INTx / DEASSERT\_INTx messages per the *PCI Express Base Specification*. In MR, these messages are queued and ordered within a VH.

#### 2.5.1.1. INTx Device Processing

In PCIe, INTx messages are emitted by Devices. An ASSERT\_INTx message is emitted when an interrupt condition is signaled. A DEASSERT\_INTx message is emitted when the interrupt condition has been satisfied.

In MR, INTx messages are emitted within a VH. If a function within a VH signals or satisfies an interrupt, ASSERT\_INTx / DEASSERT\_INTx messages are emitted within that VH. These messages are unrelated to INTx messages issued in any other VH.

#### *2.5.1.2. INTx Switch Processing*

In PCIe, INTx messages are processed by Switches. Each downstream Switch Port tracks an internal INTx wire for each of INTA/B/C/D. These INTx wires are combined into four INTx wires at the upstream Switch Port. Transitions of the combined INTx wires trigger sending of INTx messages out the upstream Switch Port.

Similarly, in MR, INTx messages are processed by Virtual Switches. Each virtual downstream Switch Port tracks an internal INTx wire for each of INTA/B/C/D. The INTx wires are combined into four INTx wires at the virtual upstream Switch Port. Transitions of the combined INTx wires trigger sending of INTx messages out the virtual upstream Switch Port.

Switch reconfiguration will also affect INTx. For example, when a Virtual Device is unmapped from a Virtual Switch, the virtual downstream Switch Port sees DL\_Down. This virtual downstream Switch Port follows PCIe rules by deasserting the internal INTx wires. If this results in a transition of the combined INTx wires a DEASSERT\_INTx message is sent out the virtual upstream Switch Port.

#### *2.5.1.3. INTx Root Port Processing*

In PCIe, the Host Processor tracks the INTx assert/deassert state. If INTx is asserted, and software has not masked processing, an interrupt is sent to host processor.

In MR with a PCIe Root Port, this is unchanged. By the time the INTx message is seen by the Root Port, the TLP Prefix has been dropped and the message is indistinguishable from non-MR usage.

In MR within an MR Root Port, the INTx messages are tagged with a VH and the RP must track independent INTx assert/deassert state for each VH.

### 2.5.2. PME Turn Off Processing

PCIe switches must perform a PME Turn Off scoreboard function. MR switches must do so as well within each VS. See Section 7.7 for details.

### 2.5.3. PM\_PME Processing

MR Switches must convert PM\_PME messages into Beacon / WAKE# indications in certain situations. See Section 7.6.1 for details.

## 2.6. Miscellaneous Changes

There are certain secondary effects caused by other MR Protocol changes.

- ❑ TLPs in a retry buffer are slightly bigger since it includes a TLP Prefix. This has a minor ripple effect (e.g. Ack/Nak timer values may need minor adjustments)
- ❑ The TLP Prefix is queued with the TLP. This increases the amount of buffering required for the TLP Header.

## 2.7. Miscellaneous Non-Changes

The following items are not affected by Multi-Root operations.

- ❑ Locked transactions are processed by MR Switches within each VS. Locked transactions in one VS will not affect another VS.
- ❑ Port arbitration occurs within each VS. Software in a VH may control Port arbitration using the optional VC Extended Capability in each bridge of the VS.
- ❑ Function arbitration occurs within each VH. Software in a VH may control Function arbitration using the optional MFVC Extended Capability in Function 0 of each VH.
- ❑ Values reported by the Power Budgeting Capability reflect the entire component. No attempt is made to split reported power consumption between VHs.

## 3. Initialization and Resource Allocation

### 3.1. MR Topology Initialization

MR Fabrics consist of a collection of the following components interconnected in a mesh topology:

- ❑ one or more MRA Switches
- ❑ one or more PCIM enabled Switch Management Ports
- ❑ zero or more non-PCIM enabled RPs
- ❑ zero or more Devices – MRA, SR, PCIe Device<sup>2</sup>

Unlike conventional PCI Express, MR topologies need not be a tree. MR Topologies can be an arbitrary mesh that may contain loops. Each VH sees a tree structure consisting of a subset of the overall MR Topology.

Because MR Topologies are not a tree, they have no single notion of link upstream and downstream directions. Links have a *Physical* direction (used in the Physical layer Link training process). In addition, each VH using a Link has a *Logical* direction which may differ from the Link's Physical direction.

There are a number of steps or phases used to initialize and configure the above components into a MR Topology. These steps are:

1. Initial State after Reset
2. PCIM Location Policy Decision
3. Topology Discovery
4. Component Discovery
5. Mapping Policy Decision
6. Mapping Implementation
7. Virtual Hierarchy Enumeration

Each step is described in more detail below.

---

<sup>2</sup> Useful MR Topologies will eventually have at least one device, but this is not strictly required. In particular, no devices might initially be present with devices being added later using Hot-Add operations.

### 3.1.1. Initial State after Fundamental Reset

MRA Switch Ports are configured into one of two Initial Port Types: *PCIM Capable Switch Port* and *Non-PCIM Capable Switch Port*. This configuration is set after Fundamental Reset using a Vendor Specific mechanism. The configuration mechanism can change Port type assignments but the results of such a change are not visible until the next Fundamental Reset.

The initial configuration of every MR Switch must have at least one Switch Management Port. This port can be either a PCIe PCIM Capable Switch Port or a Vendor Specific non-Pcie Switch Management Port.

The initial configuration of every MR Switch may have any number of Non-PCIM Ports (including zero).

Initial Port Type is indicated by the setting of a few key characteristics of a Port. These characteristics are configurable so that MR-PCIM software can change PCIe Port behavior from the initial settings. Port Type configurability for non-Pcie Switch Management Ports is Vendor Specific.

Implementation Note: Vendor specific settings could be used to configure an MR Switch so it operates as an SR Switch. The SR upstream Port would be a PCIM Capable Switch Port. SR downstream Ports would be Non-PCIM Ports with their Link\_Direction set as downstream and their Port VH 0 mapped to some downstream P2P Bridge of the VS associated with the upstream Port.

#### 3.1.1.1. PCIM Capable Switch Ports

A PCIM Capable Switch Port *i* is defined as a Switch Port with the following characteristics:

- ❑ Port[i].Link\_Direction indicates an upstream Port.
- ❑ Some VS[j] has mapped {Port[i], Port VH 0} to its upstream bridge.
- ❑ The upstream P2P Bridge Configuration Header for VH 0, VS[j] has a full MR-IOV Capability.
- ❑ The bit *j* of the VS Authorization Bitmap is Set.
- ❑ The Management VS value is Vendor Specific. It could be *j* or it could be some other VS.
- ❑ VS[j] contains sufficient Enabled Downstream P2P Bridges to access all Ports of the MR Switch.

These settings ensure that MR-PCIM could manage the Switch using this Port. MR-PCIM is not required to be present on this Port. The RP running MR-PCIM could be directly attached to this Port or could be connected indirectly using one or more MR or SR Switches.

This Port is authorized. In particular, any software using it will be allowed to manage the Switch. Software can later de-authorize the Port if desired.

Note: Since a VS has a single upstream bridge, these rules imply that every Potential PCIM port will be associated with a distinct VS.

---

#### Implementation Note:

PCIM Capable Switch Ports need not be full width “expensive” Ports. They could be x1 ports intended just to support management.

---

### 3.1.1.2. *Non-PCIM Capable Switch Ports*

A Non-PCIM Capable Switch Port  $j$  is defined as a Switch Port with the following characteristics:

- ❑ Port[j].Link\_Direction is Vendor Specific.
- ❑ No Authorized VS has mapped {Port[j], any Port VH} to its upstream bridge
- ❑ Mapping of Port[j] into bridges in VSs is otherwise Vendor Specific.
  - Any Port VH of Port[j] may be mapped to any downstream bridge of any VS.
  - Any Port VH of Port[j] may be mapped to the upstream bridge of any non-authorized VS. This upstream bridge may or may not have an MR-IOV capability in its Type 1 Configuration header. If it has one, only selected fields in the Type 1 header are operational and none of the MR-IOV tables located in memory are visible. See <REF> for details.
  - Port VHs of Port[j] need not be mapped into any VS.

These settings ensure that Initial MR-PCIM will never be present on this Port. This Port can be directly or indirectly via Switches connected to Devices, Root Ports or Bridges.

This Port is not authorized. Attempts by software attached to this Port to configure the MR Switch will fail unless the port is later Authorized

### 3.1.1.3. *Non-PCIe Switch Management Ports*

A non-PCIe port may also be used to manage MR Topologies. These ports consist of Vendor Specific hardware that appears to a MR Switch as an upstream Port. Such Ports have a Port Table entry with the Non-PCIe bit set.

This Vendor Specific hardware allows MR-PCIM to issue and respond to the subset of PCIe transactions needed to manage the MR Topology.

- ❑ Such hardware must allow MR-PCIM to issue:
  - Configuration Read and Write Requests (Type 0 and Type 1) of sizes 1, 2 or 4 bytes, naturally aligned.
  - Memory Space Read and Write Requests (32 bit and 64 bit addressing) of sizes 1, 2, 4 and 8 bytes, naturally aligned.
  - Message Transactions normally issued by Root Ports (e.g. PME Turn Off messages).
- ❑ Such hardware must allow MR-PCIM to respond to:



- Completions related to the above (including completion status and support for Completion Timeout).
- Posted Memory Write transactions for MSI Interrupts
- Message Transactions (e.g. INTx, PME\_TO\_Ack, PM\_PME, Errors, ...)

This is the minimum support needed to configure the MR Topology. Additional support may be required if MR-PCIM uses the MR Topology for general I/O (e.g. Logging). Additional support may also be required if Vendor Specific device management software also needs to use this port.

A Non-PCIe Switch Management Port *i* is defined as a Port with the following characteristics:

- ❑ Port[*i*].Link\_Direction indicates an upstream Port.
- ❑ The Port operates in Base PCIe mode (the link can not be MR Enabled).
- ❑ Some VS[*j*] has mapped {Port[*i*], Port VH 0} to its upstream bridge. This mapping may be fixed.
- ❑ The upstream P2P Bridge Configuration Header for VH 0, VS[*j*] has a full MR-IOV Capability.
- ❑ The bit *j* of the VS Authorization Bitmap is Set. This bit may be read only.
- ❑ The Management VS value is Vendor Specific. It could be *j* or it could be some other VS.
- ❑ VS[*j*] contains enough Enabled Downstream P2P Bridges to access all Ports of the MR Switch.

These settings ensure that MR-PCIM could manage the Switch\_using this Port. MR-PCIM is not required to be present on this Port.

This Port is authorized. In particular, any software using it will be allowed to manage the Switch.

Note: Since a VS has a single upstream bridge, these rules imply that every Non-PCIe Switch Management Port will be associated with a distinct VS.

3.1.1.4. Initial State Example

Figure 3-1 shows an example MR Topology. Both Switches are MRA Switches. The three green Root Ports labeled RP 0, RP 1 and RP 2 are capable of running MR-PCIM software and configuring the topology. The blue RP 3 is not permitted to run MR-PCIM. Devices are a mixture of Base PCIe, SR Aware and MR Aware Devices.

Deleted: Figure 3-1

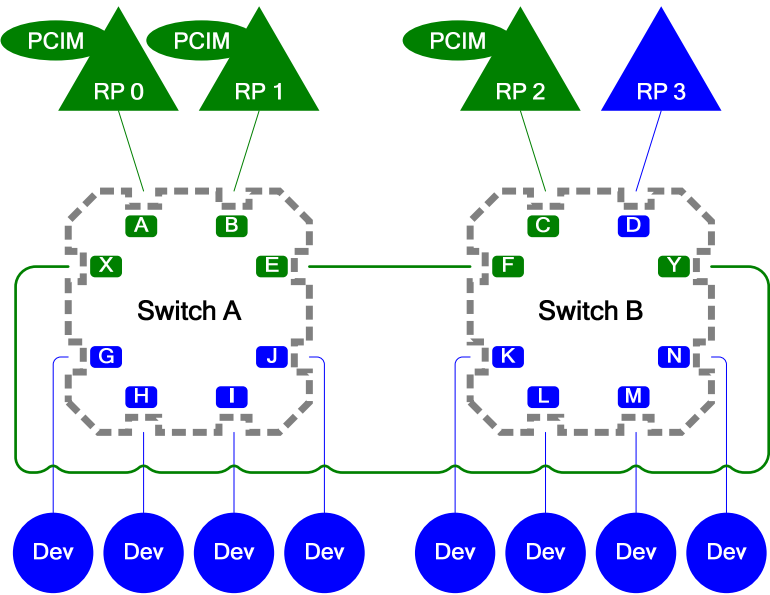


Figure 3-1: Example MR Topology

In this example, assignments are show in Table 3-1.

Deleted: Table 3-1

Table 3-1: Port Types – Example MR Topology

Port	Initial Port Type	Reason
A, B, C	PCIM Capable Switch Port	After Reset, RP 0, RP 1 or RP 2 are allowed to run MR-PCIM
D	Non-PCIM Capable Switch Port	After Reset, RP 3 is not allowed to run MR-PCIM
E, X	PCIM Capable Switch Port	If RP 2 runs MR-PCIM, this setting allows it to manage Switch A after Reset
F, Y	PCIM Capable Switch Port	If either RP 0 or RP 1 runs MR-PCIM, this setting allows it to manage Switch B after Reset
G, H, I, J	Non-PCIM Capable	Devices never run MR-PCIM

	Switch Port	
K, L, M, N	Non-PCIM Capable Switch Port	Devices never run MR-PCIM

### 3.1.2. Initial MR-PCIM Location Policy

After Fundamental Reset, a Vendor Specific mechanism selects the Initial MR-PCIM location.

Only the selected system is allowed to access the MR topology; a Vendor Specific mechanism prevents other RPs and Non-PCIE Switch Management Ports from accessing the MR topology.

The Initial MR-PCIM software must be able to manage the entire MR Topology. To do so, it must be connected to a PCIM Capable Switch Port on every MR Switch. This connection can be direct or indirect using additional Switches.

#### *3.1.2.1. Initial MR-PCIM Location Example*

Continuing with the example from Section 3.1.1.4, assume that RP 0 is chosen for the Initial MR-PCIM. Vendor Specific mechanisms are used to prevent RP 1 and RP 2 from accessing the MR Topology. For example, the affected processors might be powered off or held in reset.

### 3.1.3. Topology Discovery

The selected Initial MR-PCIM can connect to the first MR Switch in a variety of ways:

- ❑ Base PCIe RP directly connected to a PCIM Capable Switch Port on the MR Switch. The link trains in Base PCIe mode. Only VH 0 of this Port will be used.
- ❑ MR enabled RP directly connected to a PCIM Capable Switch Port on the MR Switch. The link trains in MR Enabled mode. VH 0 of this Port will be used by MR-PCIM to manage this switch.
- ❑ Vendor Specific non-PCIe interface connected to a Non-PCIe Switch Management Port on the MR Switch.

The in each of these cases, the upstream P2P Bridge first seen by MR-PCIM contains the MR-IOV capability and is mapped to VH 0 of an authorized VS. MR-PCIM uses this capability to configure the Switch.

MR-PCIM invents a unique number for the Switch and writes it to the MR Switch Number field. This number is by used MR-PCIM to detect topology loops during the enumeration process.

MR-PCIM then uses the Port table to enable additional links and to manage their Link\_Direction. MR-PCIM might use Vendor Specific of knowledge of the topology to aid in this process if, for example, a system is using Link\_Direction to prevent link training and thus hold off some RPs from seeing the hierarchy until it is configured.

For each Port that trains as a downstream Port, MR-PCIM will examine the Link Partner Training Status fields in the Port Table. For each Link Partner that is an Authorized Port on an MR Switch, MR-PCIM will determine whether the Port is mapped into the MR-PCIM VS and if necessary map the Port into an unused downstream bridge. It will then establish a Bus Number range for the downstream bridge so that allow PCIe Configuration transactions can be issued to the Link Partner.

MR-PCIM then probes the configuration header of the MR Switch attached to the Port.

- ❑ If the component is a “new” MR Switch because the Switch’s MR Switch Number field has a number not assigned by MR-PCIM in this enumeration cycle, the enumeration process repeats to configure the new Switch.
- ❑ If the component is an “old” MR Switch that MR-PCIM has seen before, the MR Switch Number and connection information is noted but further enumeration is not needed via this connection. Note that enumeration of the “old” Switch may not be complete, but it will be completed using other links into the Switch.

### 3.1.3.1. MR-PCIM Topology Discovery Example

Figure 3-2 expands on the example shown in Figure 3-1, adding possible initial link directions.

Deleted: Figure 3-2

Deleted: Figure 3-1

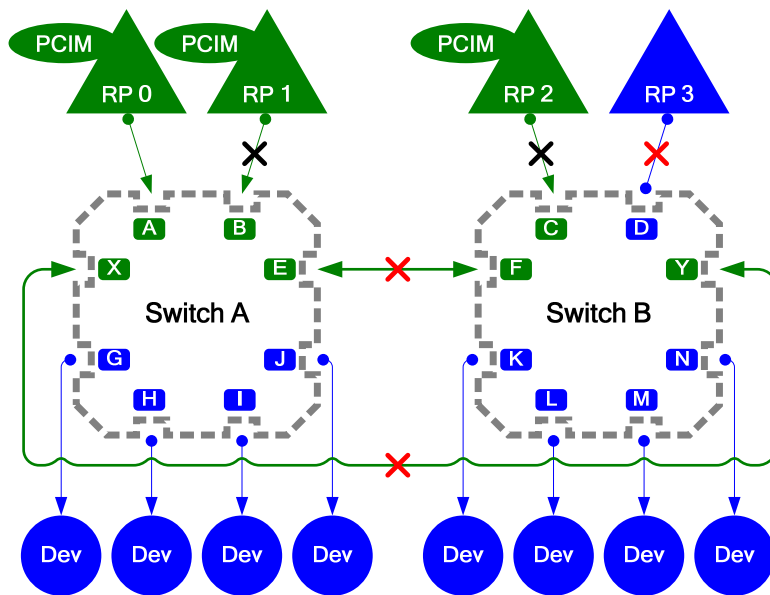


Figure 3-2: Example MR Topology with Initial Link Directions

The Links between Ports E and F and between Ports X and Y do not train since each Link consists of two upstream Ports. Since RP 0 was chosen as the location of the Initial MR-PCIM the link

between RP 0 and Switch A trains. Vendor Specific mechanisms are used to prevent RP 1 and RP 2 from starting (links may train but no transactions will be generated from these RPs). The link between RP 3 and Port D can't train since it consists of two downstream Ports.

MR-PCIM software operating in RP 0 could proceed as follows:

1. Reads Type 1 Configuration Header at Port A, detects MR-IOV capability indicating an MR Switch.
2. Configures BAR registers in Switch A so that the Port Table can be examined.
3. Assigns MR Switch Number 42 to Switch A.
4. Detects that Port A, VH 0 was mapped to VS n.
5. Notices link attached to Port E detected something present but did not train. Switches Link\_Direction for link E to downstream thus allowing the E to F link to train.
6. Notices link attached to Port X detected something present but did not train. Switches Link\_Direction for link X to downstream thus allowing the X to Y link to train.
7. Notices that Links G, H, I and J trained as downstream. Each Port's Link Partner Training Status indicates no MR Switch is connected so no additional enumeration is needed at this time.
8. Notices that Port E is connected to an Authorized MR Switch. If needed, maps Port E to some downstream bridge of VS n of Switch A (it might already be mapped). Using this downstream bridge Switch B is enumerated.
9. Reads the Type 1 Configuration Header of Port F, detects MR-IOV capability indicating an MR Switch.
10. Configures BAR registers in Switch B so that the Port Table can be examined.
11. Assigns MR Switch Number 86 to Switch B.
12. Detects that Port F, VH 0 was mapped to VS m.
13. Notices that Links K, L, M and N trained as downstream. Each Port's Link Partner Training Status indicates no MR Switch is connected so no additional enumeration is needed at this time.
14. Notices link attached to Port D detected something present but did not train. Switches Link\_Direction for link D to upstream thus allowing the D to RP 3 link to train. This step might be delayed until later if preventing link training was the Vendor Specific mechanism used hold off RP 3 from enumerating the topology.
15. Notices that Port X is connected to an Authorized MR Switch. If needed, maps Port X to some downstream bridge of VS n of Switch A (it might already be mapped). Using this downstream bridge Switch B is enumerated. Reads the Type 1 Configuration Header of Port Y, detects MR-IOV capability indicating an MR Switch. The previously assigned MR Switch Number of 86 indicating that Port X is a second path to Switch B. Since Switch B has been (or is still being) enumerated using Port E, no further enumeration is needed using Port X.

This is an example; other enumeration orderings are also valid. Note that the link directions of the E to F and X to Y links depend on the chosen ordering. If, for example, instead of changing the

Link\_Direction of Port X in Step 6 the Link\_Direction of Port Y were changed later during the enumeration of Switch B, the X to Y link would train in the opposite direction.

Note: MR-PCIM assigns MR Switch registers locations in PCIe Memory Space during this discovery process. These assignments may change in subsequent software.

### 3.1.4. Component Discovery

Topology Discovery allows MR-PCIM to know what MR Switches exist and how they are interconnected. Component Discovery is now used to locate MR Devices and non-MR Components.

This process examines config space of every Component. This information is gathered to drive Switch and Device Configuration Policy Decisions.

Information gathered from MR Devices includes:

- ☐ MaxVH
- ☐ Function number(s) of all BFs
- ☐ VF Mapping Supported / VF Migration Supported / VF MVF Region

Information gathered from MR Switches includes:

- ☐ MaxVH for each Port
- ☐ Number of VSS / Number of Bridges for each VS

Information gathered from every Link includes:

- ☐ Link Width and Speed

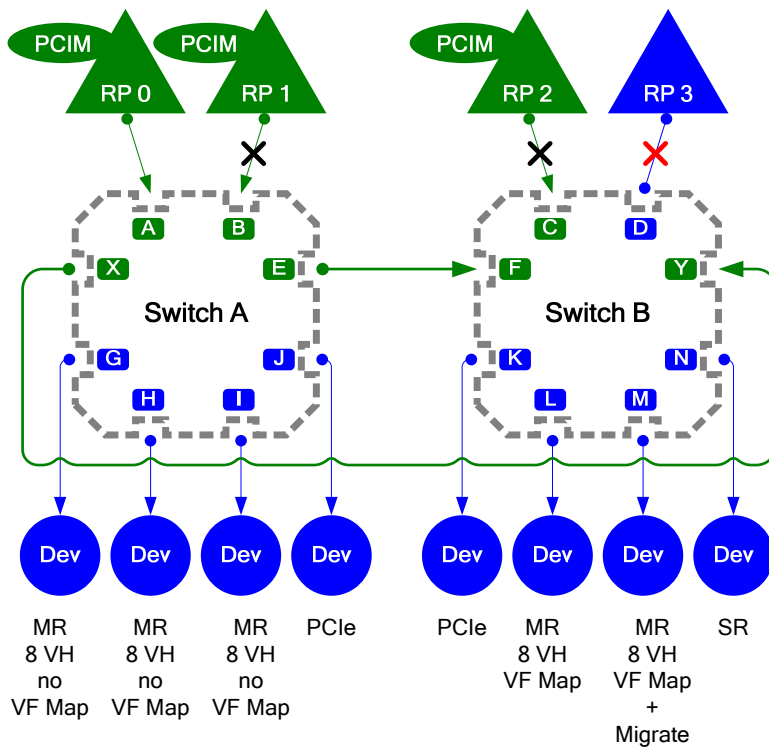
Information may not be available for Root Ports (MR or PCIe). The Vendor Specific mechanisms used to hold off transactions might also prevent the link from training so the Link Partner Training Status may not yet be meaningful.

### 3.1.4.1. Component Discovery Example

Figure 3-3 expands on the example shown in Figure 3-2, adding Device characteristics.

Deleted: Figure 3-3

Deleted: Figure 3-2



**Figure 3-3: Example MR Topology with Component Discovery Details**

This example assumes:

- ❑ Single Function MR Devices attached to Ports G, H and I
  - Each Device supports 8 VHs
  - None of the Devices support VF Mapping or VF Migration
- ❑ Base PCIe Devices attached to Ports J and K
- ❑ Single-Function MR Device attached to Port L
  - The Device supports 8 VHs
  - VF Mapping is supported with 32 LVFs and 32 MVFs

- VF Migration is not supported
- ❑ Single Function MR Device attached to Port M
  - The Device supports 8 VHs
  - Both VF Mapping and VF Migration are supported with 32 LVFs and 30 MVFs
  - VF and Function MVF Regions fully overlap
- ❑ Single Root Device attached to Port N

### 3.1.5. VH and VF Mapping Policy

Out-of-Scope mechanisms are used to decide what portions of the MR Topology should be assigned to each VH.

#### 3.1.5.1. Example VH and VF Mapping Policy

Table 3-2 expands on the example shown in Figure 3-3 adding VF and VF Mapping Policy Decision information.

Deleted: Table 3-2

Deleted: Figure 3-3

**Table 3-2: Example MR Topology VH and VF Mapping Policy**

VH	Authorized	VH Mapping	VF Mapping
RP 0	Yes (MR-PCIM)	VS in Switch A and B Uses E to F Inter-Switch Link VH in Devices G, H, I, L and M	Device L: 16 VFs Device M: 8 VFs (2 unpopulated)
RP 1	No	VS in Switch A VH in Devices G and H Device J	
RP 2	Yes (Backup MR-PCIM)	VS in Switch A and B Uses F to E Inter-Switch Link VH in Devices G, H, I, L and M Device K	Device L: 8 VFs Device M: 8 VFs (2 unpopulated)
RP 3	No	VS in Switch A and B Uses Y to X Inter-Switch Link VH in Devices G, H, L and M Two VHs in Device I Device N	Device L: 8 VFs Device M: 8 VFs (2 unpopulated)



### 3.1.6. VH and VF Mapping Implementation

Once the assignment policy is decided, MR-PCIM configures the Components to implement it. This involves writing the various tables in MR Switches and Devices. These tables are described in Sections 4.2 and 4.3.

All configuration activity occurs in the MR-PCIM VH. The order software uses for this initial configuration writes is unspecified. Configuration and memory requests originate exclusively from MR-PCIM.

#### 3.1.6.1. Example Topology: Switch Implementation

The view from each Root Port is shown in [Figure 3-4](#) through [Figure 3-7](#) below. Switch VS Table Programming to implement this is shown in [Table 3-3](#) and [Table 3-4](#).

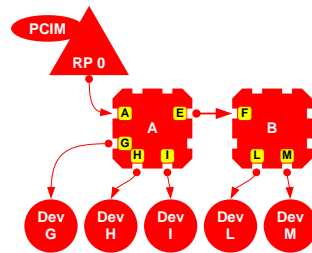


Figure 3-4: Example MR Topology: RP 0 View

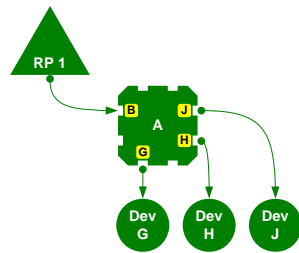


Figure 3-5: Example MR Topology: RP 1 View

Deleted: Figure 3-4

Deleted: Figure 3-7

Deleted: Table 3-3

Deleted: Table 3-4

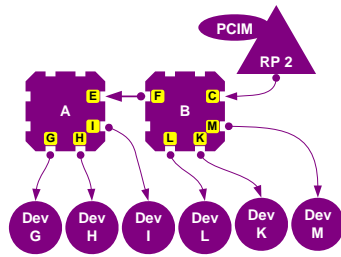


Figure 3-6: Example MR Topology: RP 2 View

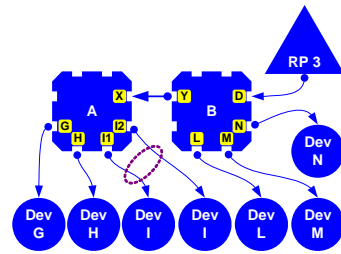


Figure 3-7: Example MR Topology: RP 3 View

**Table 3-3: Example Topology: Switch A VS Bridge Table Contents**

Slot	Root	VS	Bridge	Enabled	Mapped	Port	Port VHN
0	RP 0	VS 0	Upstream	Yes	Yes	A	VH 0
1	RP 0	VS 0	Downstream 0	Yes	Yes	G	VH 0
2	RP 0	VS 0	Downstream 1	Yes	Yes	H	VH 0
3	RP 0	VS 0	Downstream 2	Yes	Yes	I	VH 0
4	RP 0	VS 0	Downstream 3	Yes	Yes	E	VH 0
5		VS 0	Downstream 4	No			
6		VS 0	Downstream 5	No			
7		VS 0	Downstream 6	No			
8		VS 0	Downstream 7	No			
9	RP 1	VS 1	Upstream	Yes	Yes	B	VH 0 (PCIe)
10	RP 1	VS 1	Downstream 0	Yes	Yes	G	VH 2
11	RP 1	VS 1	Downstream 1	Yes	No		
12	RP 1	VS 1	Downstream 2	Yes	Yes	H	VH 1
13	RP 1	VS 1	Downstream 3	Yes	Yes	J	VH 0 (PCIe)
14	RP 1	VS 1	Downstream 4	Yes	No		
15	RP 1	VS 1	Downstream 5	Yes	No		
16	RP 1	VS 1	Downstream 6	Yes	No		
17	RP 1	VS 1	Downstream 7	Yes	No		
18	RP 2	VS 2	Upstream	Yes	Yes	D	VH 0 (PCIe)
19	RP 2	VS 2	Downstream 0	Yes	Yes	L	VH 2
20	RP 2	VS 2	Downstream 1	Yes	Yes	M	VH 2
21	RP 2	VS 2	Downstream 2	Yes	Yes	N	VH 0 (PCIe)
22	RP 2	VS 2	Downstream 3	Yes	Yes	Y	VH 0
23	RP 2	VS 2	Downstream 4	Yes	No		
24	RP 2	VS 2	Downstream 5	Yes			
25		VS 2	Downstream 6	No			
26		VS 2	Downstream 7	No			
27		VS 3	Upstream	No			
28		VS 3	Downstream 0	No			
29		VS 3	Downstream 1	No			
30		VS 3	Downstream 2	No			
31		VS 3	Downstream 3	No			
32		VS 3	Downstream 4	No			
33		VS 3	Downstream 5	No			
34		VS 3	Downstream 6	No			
35		VS 3	Downstream 7	No			

**Table 3-4: Example Topology: Switch B VS Bridge Table Contents**

Slot	Root	VS	Bridge	Enabled	Mapped	Port	Port VHN
0	RP 2	VS 0	Upstream	Yes	Yes	C	VH 0 (PCIe)
1	RP 2	VS 0	Downstream 0	Yes	Yes	F	VH 1
2	RP 2	VS 0	Downstream 1	Yes	Yes	L	VH 0
3	RP 2	VS 0	Downstream 2	Yes	Yes	M	VH 0
4	RP 2	VS 0	Downstream 3	Yes	Yes	K	VH 0 (PCIe)
5		VS 0	Downstream 4	No			
6		VS 0	Downstream 5	No			
7		VS 0	Downstream 6	No			
8		VS 0	Downstream 7	No			
9	RP 0	VS 1	Upstream	Yes	Yes	F	VH 0 (PCIe)
10	RP 0	VS 1	Downstream 0	Yes	Yes	L	VH 1
11	RP 0	VS 1	Downstream 1	Yes	Yes	M	VH 1
12	RP 0	VS 1	Downstream 2	Yes	No		
13	RP 0	VS 1	Downstream 3	Yes	No		
14	RP 0	VS 1	Downstream 4	Yes	No		
15	RP 0	VS 1	Downstream 5	Yes	No		
16	RP 0	VS 1	Downstream 6	Yes	No		
17	RP 0	VS 1	Downstream 7	Yes	No		
18	RP 2	VS 2	Upstream	Yes	Yes	E	VH 1
19	RP 2	VS 2	Downstream 0	Yes	Yes	G	VH 1
20	RP 2	VS 2	Downstream 1	Yes	No		
21	RP 2	VS 2	Downstream 2	Yes	Yes	H	VH 2
22	RP 2	VS 2	Downstream 3	Yes	Yes	I	VH 1
23		VS 2	Downstream 4	No			
24		VS 2	Downstream 5	No			
25		VS 2	Downstream 6	No			
26		VS 2	Downstream 7	No			
27	RP 3	VS 3	Upstream	Yes	Yes	X	VH 0
28	RP 3	VS 3	Downstream 0	Yes	Yes	G	VH 3
29	RP 3	VS 3	Downstream 1	Yes	Yes	H	VH 3
30	RP 3	VS 3	Downstream 2	Yes	Yes	I	VH 3 <sup>3</sup>
31	RP 3	VS 3	Downstream 3	Yes	Yes	I	VH 2 <sup>4</sup>
32		VS 3	Downstream 4	No			

<sup>3</sup> Port I1 in [Figure 3-7](#). Two PFs of Device I are assigned to RP 3. As far as RP 3 is concerned, these are independent Devices.

<sup>4</sup> Port I2 in [Figure 3-7](#).

Deleted: Figure 3-7

Deleted: Figure 3-7

33		VS 3	Downstream 5	No			
34		VS 3	Downstream 6	No			
35		VS 3	Downstream 7	No			

### 3.1.6.2. Example Topology: Device Implementation

In addition to Switch mapping, MR Devices need configuration.

Devices G and H are both assigned NumVHs of 4. No additional configuration is necessary. Switch configuration established the mapping between VH numbers and RPs.

Device I is assigned NumVHs of 6. No additional configuration is necessary. VH 4 and VH 5 of the Device could be dynamically assigned to unmapped switch ports associated with RP 0, RP 1 or RP 2 (RP 3 has no unmapped switch ports on Switch A).

Device L is assigned NumVHs of 3. VF Mapping is shown in [Figure 3-8](#).

Deleted: Figure 3-8

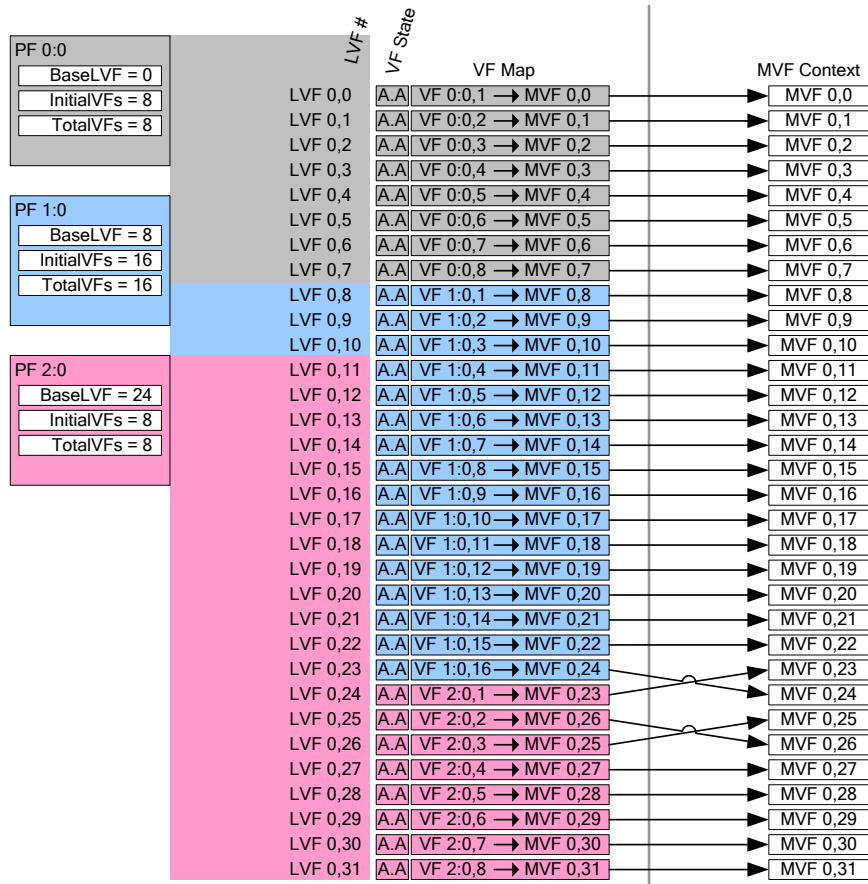


Figure 3-8: Example MR Topology: Device L PF / VF Mapping

Device M is assigned NumVHs of 3. VF Mapping is shown in [Figure 3-9](#). VF Migration Capable is Set.

Deleted: Figure 3-9

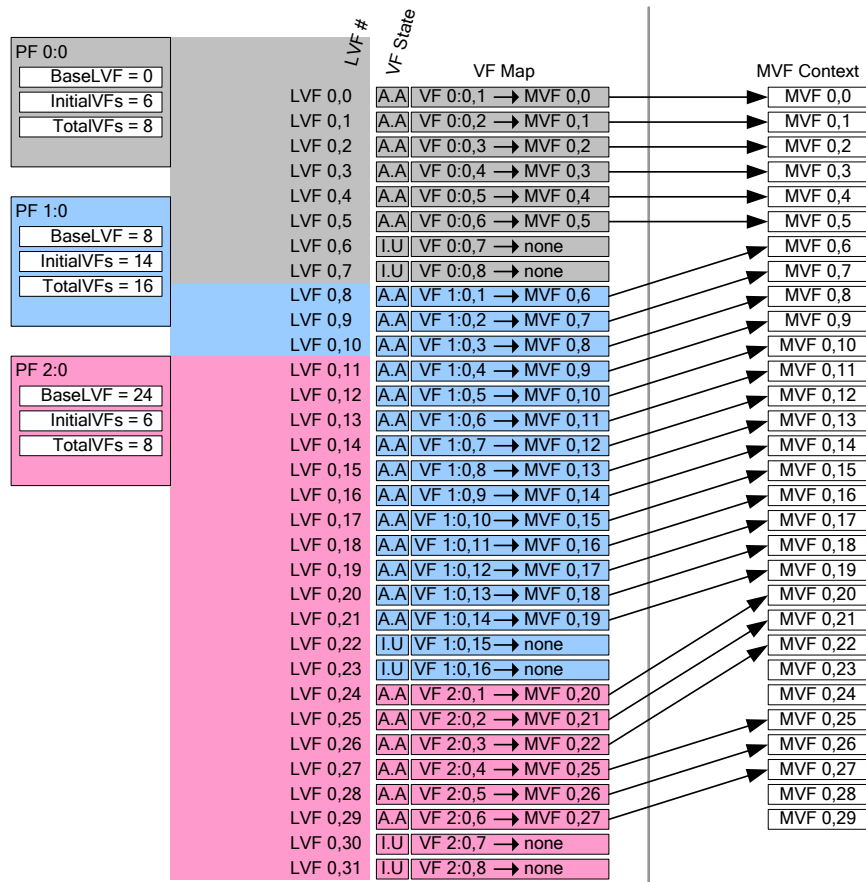


Figure 3-9: Example MR Topology: Device M VF Mapping

### 3.1.7. MR-PCIM Failover

There is a single *active* MR-PCIM in a topology. This MR-PCIM receives all “route to PCIM” errors and events. The VS associated with this MR-PCIM is programmed into each Switch’s Management VS register.

There may be multiple Authorized VSs. Software running in these VSs is allowed to view and change the Switch MR-IOV configuration. To avoid confusion, coordination is needed between such software but is not specified by this specification.

A “backup” MR-PCIM operating in any Authorized VS may become the active MR-PCIM simply by changing the Management VS register in the affected Switch(s). This feature can support failover from one MR-PCIM to a backup MR-PCIM. The Suppress Reset Propagation feature of a VS can be used to prevent a reset due to the failure of one MR-PCIM resetting state needed to continue operation using a backup MR-PCIM.

Mechanisms used to detect MR-PCIM failure, to select the new MR-PCIM location, to initiate the failover, etc. are undefined by this specification.

TLPs targeting MR-PCIM are regular PCIe TLPs within the appropriate VH. Such TLPs are forwarded based on the switch configuration when they are received. Old TLPs are not re-routed due to switch reconfiguration or change in VS Authorization.

## 3.2. MR Device Initialization

After Conventional Reset, MR Devices negotiate to use the MR Link protocol. If this negotiation is successful, the Device appears as VH 0 and contains one or more Base Functions (BFs) and zero or more PFs or Non-IOV Functions. Each BF contains an MR-IOV Capability in the Type 0 Configuration Header.

Additional VHs beyond VH0 are enabled using the MR-IOV Capability.

Every BF is associated with a single PF or Non-IOV Function in each non-zero VH.

Every BF is optionally associated with a single PF or Non-IOV Function in VH0.

Every PF (in any VH) and every Non-IOV Function (in any non-zero VH) is associated with a single BF.

Attached to each PF in each VH is an optional collection of Virtual Functions also in that VH. These are described in the SR-IOV Specification."

PFs, BFs and VFs are designated as follows:

**BF** indicates the Base Function. It does not have a number since each Device only has one BF.

**PF  $f$**  indicates a PF at Function number  $f$  ( $f$  must be between 0 and 255). This nomenclature is used for SR systems or for the SR view of a single MR VH.

**PF  $h:f$**  indicates a PF within an MR system at Function number  $f$  in **VH  $h$**  ( $h$  must be between 0 and the maximum VH number in use on the link).

**VF  $f,s$**  indicates VF number  $s$  attached to PF number  $f$  ( $s$  must be between 1 and the number of VFs in use for PF  $f$ ). This nomenclature is used for SR systems or for the SR view inside a single VH.

**VF  $h:f,s$**  indicates VF number  $s$  attached to PF number  $f$  in **VH  $h$** .



In addition, the optional VF Mapping and VF Migration features use the terms Logical Virtual Function (LVF) and Mission Virtual Function (MVF). These are similarly designated as follows:

**LVF  $f$  $s$**  indicates LVF table slot number  $s$  attached to PF number  $f$ .

**MVF  $f$  $s$**  indicates MVF number  $s$  attached to PF number  $f$ . MVFs do not have a VH (a VH is associated the LVF that an MVF is mapped to, see below).

During topology enumeration, MR-PCIM detects MR Devices by noticing the presence of an MR-IOV Capability in PF 0's Configuration Header.

Initializing and managing a Device in MR mode involves managing four aspects of the Device.

- ❑ Configuring and enabling the VHs.
- ❑ Enabling and managing the optional MR flow control features: This involves configuring the number of Virtual Links used, configuring the number of VCs offered in each VH, configuring the {VH, VC} to VL mapping hardware, configuring any VH to VL arbitration hardware and configuring any VL to Link arbitration hardware.
- ❑ Enabling and managing the optional VF Mapping features: This involves configuring the number of LVFs offered by each PF in each VH and configuring each PF's LVF to MVF map.
- ❑ Enabling and managing the optional VF Migration features: This involves leaving "holes" in the LVF to MVH map to support migration, enabling VF Migration, responding to requests for initiate VF migration and interacting with SR-PCIM software to accomplish VF migration.

These aspects will be described separately in the following sections.

### 3.2.1. Enabling MR Operation

Initially MR Devices always negotiate to use the MR Link Protocol. If this negotiation fails, they operate in PCI Express mode. Initially, MR Devices operating in MR mode use only VH 0. VH 0 contains one or more Functions (PFs, BFs or non-IOV Functions) operating as either a PCI Express single Function or Multi-Function Device.

In MR Devices, each BF contains an MR-IOV Capability block. There are a few key registers in this Capability used in enabling MR Operation.

- ❑ The MR-IOV Capabilities register indicates which optional features are implemented by the Device.
- ❑ One of the BFs is designated the Main BF. This BF contains certain fields that apply to the entire Device. These fields are reserved in other BFs (if any). The Main BF is identified by the "Is Main BF" bit in the MR-IOV Capability. The function number associated with the Main BF is Device Specific.
- ❑ The MaxVH register in the Main BF indicates the number of VHs supported by Device hardware. The value is Vendor Specific and must not change except after Fundamental Reset of the Device.

- ❑ The NumVH register in the Main BF indicates to the Device how many VHs are going to be used by MR-PCIM. Software should set this value based on the value of MaxVH, on the number of VHs implemented at the upstream end of the link and on the number of VHs needed by the system. Once software has enabled additional VHs, the NumVH value may not be changed.
- ❑ The MR-IOV Capability of every BF contains a pointer to the Function Table. This table contains one entry for every Function associated with the BF. This table is indexed by VH number since every BF contains a single function in each VH (exception: VH 0 need not have a function so the first Function Table entry might not be used see the Function Offset field in Section 4.2.4.1)
- ❑ The Function Table entry for the Main BF contains a VC ID to VL Map. This map includes a Map Enable bit. A VH is considered Enabled when, for some VC, the VC ID to VL Map entry is Enabled, points to a VL that is Enabled, and software operating in the VH enables that VC. See section 4.2.4.4 for additional details.

### 3.2.2. Managing Flow Control

There are a number of fields used to manage flow control and VC to VL mapping.

MR Flow Control uses the same concepts as PCIe flow control. TLPs can only be sent if the transmitter has sufficient available flow control credits of the appropriate flavor. The differences from PCIe include:

In MR, VCs are replaced by Virtual Links (VLs). All Flow Control information is related to VLs not VCs. VCs continue to have meaning within a VH and serve as the mechanism used to present software operating in each VH with the collection of flow control channels and the mechanism that allows designating which TLPs should get use each flow control channel. As today, software operating in a VH controls the TC to VC map. MR-PCIM controls the subsequent map from {VH, VC} to VL. Performance characteristics of a VL are assigned by MR-PCIM.

If **PF h:f** supports more than one VC, the either optional VC Capability or optional MFVC Capability exists in the PF **h:f** Type 0 Configuration Header. The values presented are managed by MR-PCIM as follows:

- ❑ The **PF h:f** Extended VC Count value is configured by MR-PCIM. This value indicates to software operating in the VH the number of VC Resources that it can use.
- ❑ The **PF h:f** Low Priority Extended VC Count value is configured by MR-PCIM. This value indicates to software operating in the VH which VC Resources should be viewed as higher priority.
- ❑ The **PF h:f** VC to VL map is configured by MR-PCIM. This map converts VC IDs chosen by software running in the VH into the corresponding VL numbers. If more than one VC is supported, this is a full eight entry map since all VC ID values are legal This map also includes an Enable bit.

- ❑ The **PF h:f** VC Enabled fields allow MR-PCIM to determine which VC Resources are enabled by software operating in the VH. When VC Enabled changes value, the VC Config Changed interrupt is signaled to MR-PCIM.
- ❑ The **PF h:f** VC ID fields allow MR-PCIM to determine the VC ID assignments made by software operating in the VH.
- ❑ The **PF h:f** TC to VC map field allows MR-PCIM to determine the mapping between TCs and VCs made by software operating in the VH.

### 3.2.3. Managing VF Mapping

Software operating in the VH (e.g. SR-PCIM) can optionally see a collection of Virtual Functions attached to each PF. In MR, these VFs are known as Logical Virtual Functions (LVFs). MR Devices implement some number of underlying Mission Virtual Functions (MVF). These MVFs are mapped into LVFs. This mapping can be optionally controlled by MR-PCIM using VF Mapping hardware. This VF Mapping support is optional. If VF Mapping is not supported for a BF, LVFs are still mapped to MVFs but the mapping is Device Specific and is not visible or controllable by MR-PCIM.

**VF Mapping** uses a set of mapping tables controlled by management software. These tables allow software to (1) control the number of LVFs assigned to each PF, (2) specify which MVFs (if any) are assigned to each LVF and (3) detect how many VFs software operating in the VH has indicated it will use.

Figure 3-10 shows an example setup. There are 5 VHS, each with a single PF at Function 0. VH 1 through VH 4 have been assigned some VFs.

Deleted: Figure 3-10

- ❑ VFs in VH 1 have been assigned 8 LVFs and 4 MVFs. LVF 0-0 thru LVF 0-7 are associated with the 8 VFs. There are 4 VF holes (VF 1:0.5 thru VF 1:0.8 a.k.a. LVF 0-4 thru LVF 0-7) meaning that, if SR-PCIM enables VF Migration, up to 4 MVFs can be migrated in to VH 0.
- ❑ VFs in VH 2 have been assigned 4 LVFs and 4 MVFs. All LVFs are populated meaning that no migration in is possible unless an MVF first migrates out.
- ❑ VFs in VH 3 have been assigned 7 LVFs and 5 MVFs. Like VH 0, holes were left for possible VF Migration in to the VH.
- ❑ VFs in VH 4 have been assigned 3 LVFs and 3 MVFs, Like VH 1, no migration in is possible unless an MVF first migrates out.
- ❑ No VFs have been assigned to VH 5.

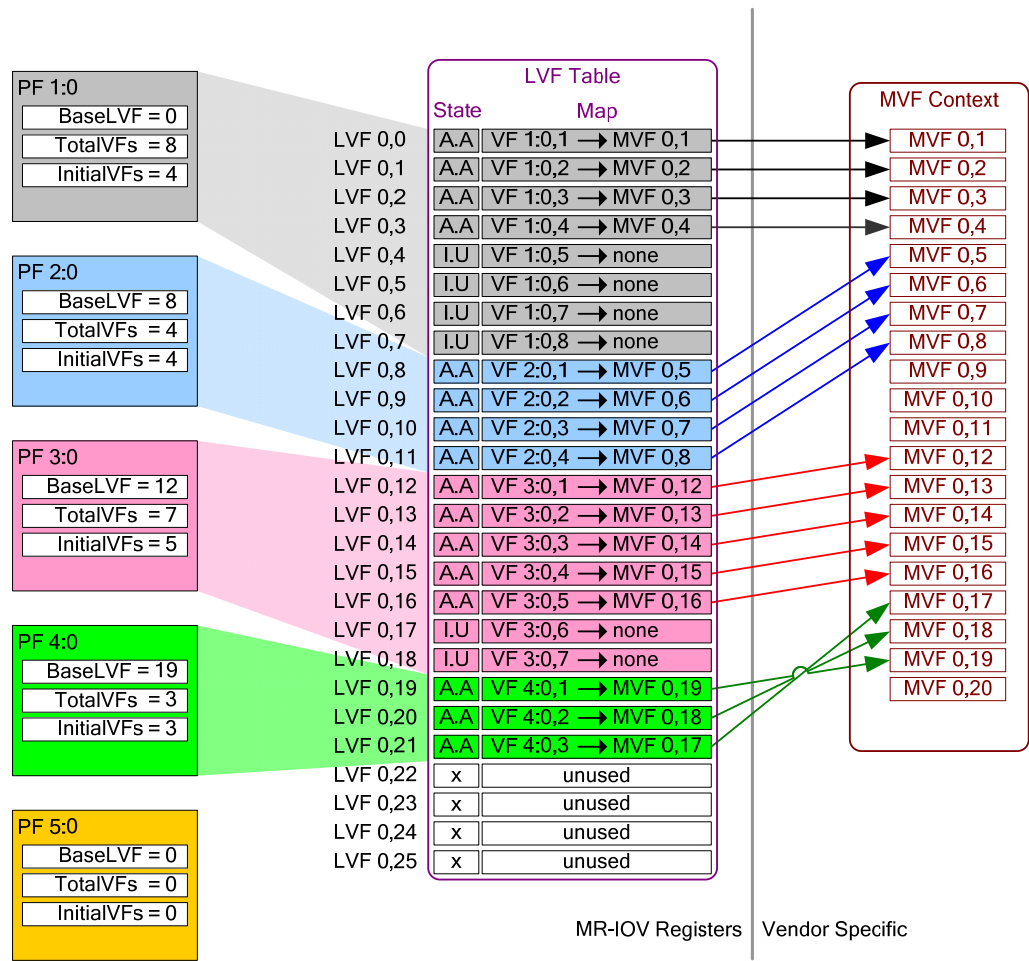


Figure 3-10: Example Mapping of VFs

The VF State table is used to coordinate migration of MVFs in / out of LVFs. If hardware supports VF Mapping but not VF Migration, the VF State table is Read Only Zero. If hardware supports VF Migration, management software must configure the VF State in the LVF Table even if the VF Migration Capable bit is Clear. The VF State values shown in Figure 3-10 are the required initial values. See Section 3.2.4 Managing VF Migration for details.

MR-PCIM manages the LVF slot allocation to PFs using the Base LVF and Total VFs registers. Base LVF indicates the first LVF slot associated with the PF. Base LVF + Total LVFs indicates the last LVF slot associated with the PF. The LVF slot designated by Base LVF contains the MVF associated with the PF. The LVF slot designated by Base LVF+1 contains the MVF associated with VF x:1 of PF x, etc. Every PF has at least one LVF slot.

Deleted: Figure 3-10

The number of populated LVFs offered to SR-PCIM is contained in InitialVFs. If SR-PCIM does not enable VF Migration, only these slots are used and any additional unpopulated slots remain unused.

Initially, MR-PCIM must populate LVFs for a given PF using the lower numbered VFs first. Holes left for migration (if any) follow the last populated LVF.

Before it enables SR-IOV operation, SR-PCIM in each VH must set the NumVFs value to indicate the number of VFs it wishes to use. The value set must be less than or equal to TotalVFs. If VF Migration not enabled by SR-PCIM, the value set must also be less than or equal to InitialVFs.

In some Devices, the Device Programming Model assumes that operations on one Function can affect another Function of the Device. In MR, this dependency manifests itself as a relationship between MVFs of some of the Device's BFs. The Function Dependency Link field in a BF indicates the presence a dependency. See section 4.2.1.2 for details. A similar dependency exists in the SR-IOV specification but there it deals with dependencies associated with VF assignment to SIs.

### 3.2.4. Managing VF Migration

In Multi-Root systems, VFs can be migrated between VHs. VF Migration does not occur in Single-Root only systems.

SR-IOV and MR-IOV support for VF Migration is *Optional*. VF Migration for VFs associated with a BF is only possible if all three of the following are true:

- ❑ The VF Migration Supported bit in the BF MR-IOV Capability is Set.
- ❑ In at least one VH (***h***), MR-PCIM software has set the VF Migration Capable bit in the Function Table entry controlling **PF *h:f***
- ❑ In that VH (***h***), SR-PCIM software controlling **PF *h:f*** has also Set the VF Migration Enabled bit.

VF Migration centers around the LVF Table:

- ❑ The VF State field manages the SR-IOV and MR-IOV combined view of the migration state of a VF. This table is used to gracefully add and remove VFs to or from VHs.
- ❑ The VF Map field maps LVFs onto MVFs. When the VF State is Inactive.Unavailable, software can write this field to implement a change.

VF migration follows the state diagram shown in [Figure 3-11](#). The state values shown are contained in the VF State field associated with the VF. State transitions indicated by solid lines are initiated by MR software by writing the new state value to the VF State field. State transitions indicated by dashed lines are typically initiated by SR-PCIM and are visible to MR-PCIM via the VF State field. The mechanisms used for this is described in the SR-IOV Specification.

Deleted: Figure 3-11

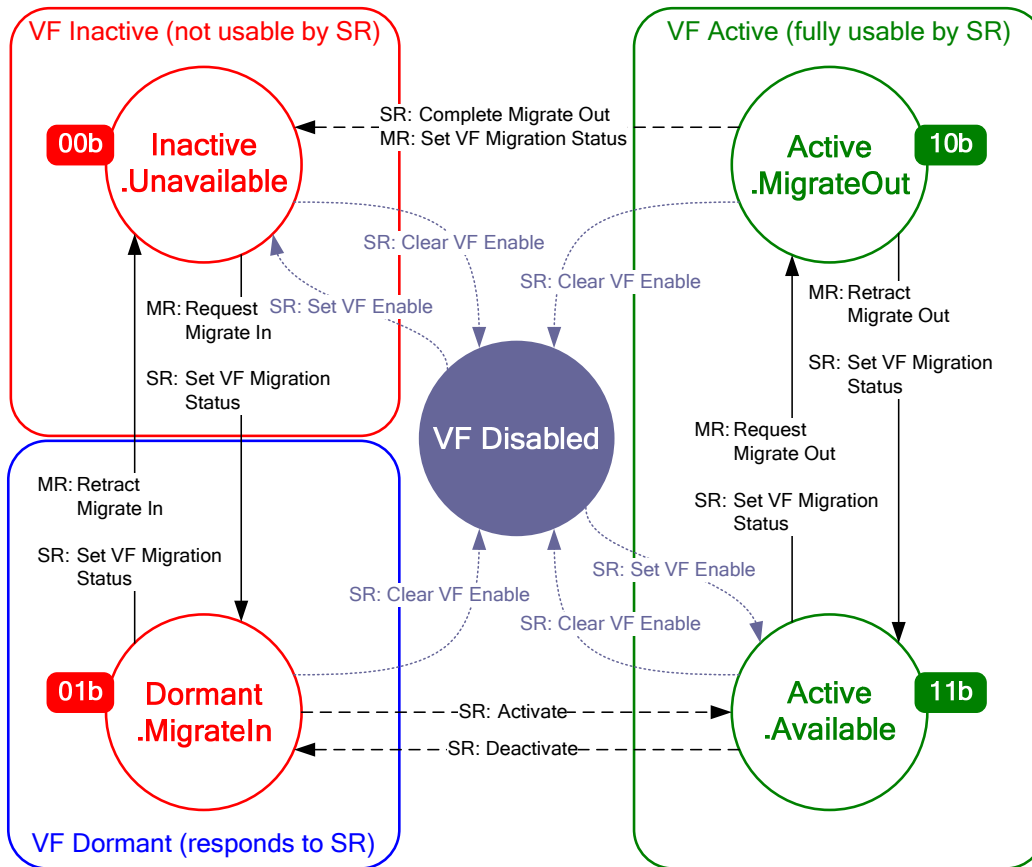


Figure 3-11: VF Migration State Diagram

The following state transitions may be initiated by MR software by writing the VF State field. Any other writes are ignored and no state transition occurs.

Table 3-5: Valid MR State Transitions for VF Migration

Current State		Written State		Meaning
00b	Inactive.Unavailable	01b	Inactive.MigrateIn	Request Migrate In
01b	Inactive.MigrateIn	00b	Inactive.Unavailable	Retract Migrate In
11b	Active.Available	10b	Active.MigrateOut	Request Migrate Out
10b	Active.MigrateOut	11b	Active.Available	Retract Migrate Out

VFs that are in the Inactive.Unavailable state are not usable by software in the VH. Configuration, IO and Memory Requests within the VH targeting the associated VF return UR. Within 100 ms of transitioning to this state, a VF must stop issuing Requests.

VFs that are in the Inactive.MigrateIn state (1) will respond to Configuration Requests issued by software running in the VH, (2) if MSE is Set, will respond to Memory requests issued by software running in the VH and (3) will not issue Requests.

The state transition from Active.MigrateOut to Inactive.Unavailable Sets the MR VF Migration Status bit. If the MR VF Migration Interrupt Enable is Set, this in turn causes an MSI to be queued to MR-PCIM. MR-PCIM software can then scan the LVF Table to determine the cause of the interrupt. Specifically, MR-PCIM is looking for the VFs that it previously placed in the Active.MigrateOut state and are now in the Inactive.Unavailable state.

State transitions with the notation “SR: Set VF Migration Status” cause similar behavior within the VH. See the SR-IOV Specification for details.

The following steps are used by MR-PCIM to migrate a VF from one VH to another.

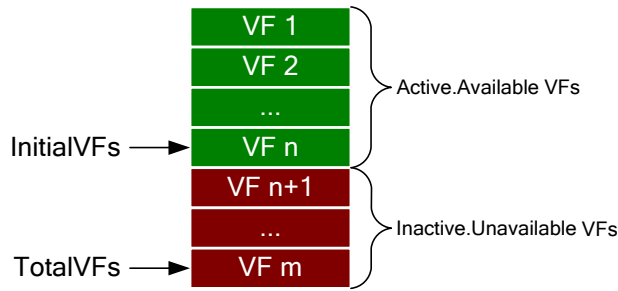
1. Request arrives from higher level software requesting that **VF *h:f:s*** be migrated to **VF *a:f:c***. For the request to be valid, the VF Table Entry associated with **VF *h:f:s*** must be in the Active.Available state and the VF Table Entry associated with **VF *a:f:c*** must be in the Inactive.Unavailable state.
2. Initiate a *Migrate Out* operation in **VH *h*** by writing the VF State entry associated with **VF *h:f:s*** to the Active.MigrateOut.
3. Wait for SR-PCIM to stop using the VF and to indicate so by transitioning the VF State to Inactive.Unavailable. This transition sets the MR VF Migration Status bit and can raise an interrupt to MR-PCIM.
4. Save the value of the VF Map entry associated with **VF *h:f:s***.
5. Set the VF Map entry associated with **VF *h:f:s*** to zero to indicate an empty slot
6. Set the VF Map entry associated with **VF *a:f:c*** the value saved in step 4.
7. Initiate a *Migrate In* operation in **VH *a*** by writing the VF Map entry associated with **VF *a:f:c*** to the Inactive.MigrateIn state.
8. At some point, SR-PCIM will transition the VF the Active.Available state and start using it.

In addition graceful migration described above, MR-PCIM can *retract* a Migrate In or Migrate Out request that it previously requested.

#### 3.2.4.1. VF Migration Initial State

Software in a VH is expecting to see the initial VF configuration shown in [Figure 3-12](#). MR-PCIM must ensure this condition by appropriate programming of the VF Migration tables.

Deleted: Figure 3-12



**Figure 3-12: Initial VF State**

Specifically, for each PF, MR-PCIM must configure the associated Function Table entry such that:

- ❑  $\text{InitialVFs} \geq 0$
- ❑  $\text{TotalVFs} \geq \text{InitialVFs}$
- ❑  $0 \leq \text{BaseLVF} + \text{TotalVFs} < \text{MaxLVF}$
- ❑ The LVF table region assigned to the PF,  $[\text{BaseLVF} .. \text{BaseLVF} + \text{TotalVFs} - 1]$ , does not overlap with the region assigned to any other PF of the BF.
- ❑ If  $\text{InitialVFs} > 0$ , all LVF entries in the range  $[\text{BaseLVF} .. \text{BaseLVF} + \text{InitialVFs} - 1]$  are in state Active.Available and are mapped to valid MVFs.
- ❑ If  $\text{InitialVFs} \neq \text{TotalVFs}$ , all LVF entries in the range  $[\text{BaseLVF} + \text{InitialVFs} .. \text{BaseLVF} + \text{TotalVFs} - 1]$  are in state Inactive.Unavailable.

#### 3.2.4.2. VF Migration Reinitialization

After a PF is Reset or when VF Enable is Cleared and then Set, a valid initial VF configuration must be re-established. The InitialVFs value may be different from an earlier initial configuration so long as the configuration meets the rules described in Section 3.2.4.1.

This process can be accomplished by hardware or software and must be completed within 1 sec (to avoid an SR software timeout resulting in the hardware being declared broken).

This process starts by adjusting InitialVFs to reflect the number of active VFs associated with the PF and then rearranging those active VFs into lower numbered VFs, keeping the same relative VF ordering.

**Implementation Note:** As described in the SR-IOV Specification, if VF Migration Enable was Set, SR software must wait 1 second after clearing VF Enable for InitialVFs and TotalVFs to become valid.

**Comment [sdg5]:** This constraint does not currently exist but will be added in the 1.0 version of SR-IOV.



### 3.3. MR Root Port Initialization

MR Root Complexes consist of one or more traditional Root Complex each with one or more Root Ports and a Vendor Specific mapping of those Root Ports into one or more MR Links.

For each MR Link, one Root Port is associated with VH 0 on that link and controls the physical link. When the Link operates in MR Mode, additional Root Ports are be associated with additional VHS of the Link.

For Links that are MR Enabled, the RP sends MRInit DLLPs as shown in Table 2-1 with the Device / Port Type indicating Root Port (0100b).

Any MR Root Complex must address the following areas. The mechanisms used are vendor specific and outside the scope of this specification.

- ❑ Method for deciding whether to enable MR operation on a given Link
- ❑ Method for associating a RP with a MR Link and a VH on that Link
- ❑ Method for arbitrating between RPs that share a MR Link
- ❑ Method for mapping {RP, VC} to VLs
- ❑ Method for enabling and disabling VLs (VL0 is always enabled, others are controlled by software)
- ❑ Method for enabling and disabling VHS (VH0 is always enabled, others are controlled by software)
- ❑ Method for controlling the number of VCs offered to each RP (could be fixed)
- ❑ Method for associating an ATPT with a RP
- ❑ Method for reporting and logging MR Errors
- ❑ Method for limiting Peer-to-Peer transactions from one RP to another RP

In addition, software running above the RC must determine how to configure and use the MR topology. This involves determining, for example, how the various VHS on each MR Link should be used, how VCs and VLs should be managed, etc.

The mechanism used to determine this is outside the scope of this specification. A variety of mechanisms can be used (in any combination) including:

- ❑ Using arbitrary out-of-band communication paths
- ❑ Using registers in Switches and/or Devices defined by this specification
- ❑ Using vendor specific registers in Switches and/or Devices
- ❑ Using the MR Switch “Link Partner” registers in MR Switches to see the values sent in MRInit DLLPs.

Formatted: Highlight

Deleted: Table 2-1

Formatted: Do not check spelling or grammar, Highlight

Formatted: Highlight

Note that the MR RP is always upstream on all VHs. This means that MR-PCIM can not, in general, access MR RP Configuration Space (Exception: MR-PCIM can access the RP it's running above and may be able to access other RPs in the same RC).

3.3.1. Example MR Root Port Topology

An example topology that includes an MR Root Complex is shown in [Figure 3-13](#).

Error! Objects cannot be created from editing field codes.

Figure 3-13: Example Topology with MR Root Complex

In this example, 5 Root Ports and Two Endpoint Functions are present in the Root Complex. The Root Ports are associated with MR Links X and Y in the following fashion.

Table 3-6: Example MR Root Topology: RP Associations

Root Port	Association
RP 0	Link X VH 0
RP 1	Link X VH 1
RP 2	Link X VH 2
RP 3	Link Y VH 0
RP 4	Link Y VH 1

If Switch X authorizes the port connected to Link X, MR-PCIM could manage Switch X via RP 0. Similarly, if Switch Y authorizes the port connected to Link Y, MR-PCIM could manage Switch Y via RP 3.

If Switches X and Y managed from the MR Root Complex and are distinct MR topologies with no connection between them, the two MR-PCIMs above RP 0 and RP 3 are independent as well. If there is a single interconnected MR topology, there must be a single MR-PCIM and it can use either RP 0 or RP 3 but not both (i.e. the Management VH must always be a tree).

Formatted: Not Highlight

Deleted: [Figure 3-13](#)

Formatted: Do not check spelling or grammar, Not Highlight

Formatted: Not Highlight

# 4

## 4. Configuration

The following sections list the configuration requirements for Base Function (BF), Physical Functions, MRA Switches and MR-PCIMs.

## 4.1. Configuration Field Summary

The following fields are used to manage MR IOV features of a Device or Switch.

**Table 4-1: MR-IOV Fields**

Field Name	Width	Type	Opt	Device Usage	Switch Usage
MSI Vector #	11	RO	Req	BF / Capability	MR-IOV Cap / Capability
MR Switch Number	16	RW	Req	n/a	MR-IOV Cap / Control
Function Dependency Link	8	RO	Req	BF / Capability	n/a
VS Interrupt Enable	1	RW	Req	n/a	MR-IOV Cap / Control
Port Interrupt Enable	1	RW	Req	n/a	MR-IOV Cap / Control
VS Interrupt Status	1	RO	Req	n/a	MR-IOV Cap / Status
Port Interrupt Status	1	RO	Req	n/a	MR-IOV Cap / Status
VS #	8	RO	Req	n/a	MR-IOV Cap / Status
VS Bridge #	8	RO	Req	n/a	MR-IOV Cap / Status
VS is Authorized	1	RO	Req	n/a	MR-IOV Cap / Status
# VS	8	RO	Req	n/a	MR-IOV Cap / Capability
# VS Bridge	8	RO	Req	n/a	MR-IOV Cap / Capability
# Port	8	RO	Req	n/a	MR-IOV Cap / Capability
VS Table Entry Size	8	RO	Req	n/a	MR-IOV Cap / Capability
VS Table Offset / BIR	32	RO	Req	n/a	MR-IOV Cap
VS Bridge Table Entry Size	8	RO	Req	n/a	MR-IOV Cap / Capability
VS Bridge Table Offset / BIR	32	RO	Req	n/a	MR-IOV Cap
Port Table Entry Size	8	RO	Req	n/a	MR-IOV Cap / Capability
Port Table Offset / BIR	32	RO	Req	n/a	MR-IOV Cap
Function Offset	16	RO	Req	BF / Capability	n/a
<b>Vendor Specific Fields</b>					
Vendor Specific Interrupt Enable Bits	4	RW	Opt	BF / Control	MR-IOV Cap / Control
Vendor Specific Interrupt Status Bits	4	RO / RW1C	Opt	BF / Status	MR-IOV Cap / Status

Field Name	Width	Type	Opt	Device Usage	Switch Usage
<b>Watchdog Timer Support</b>					
Watchdog Timer Interrupt Enable	1	RW	Req	n/a	MR-IOV Cap / Control
Watchdog Timer Interrupt Status	1	RW1C	Req	n/a	MR-IOV Cap / Status
Timer Interval 1	8	RW	Req	n/a	MR-IOV Cap / Watchdog
Timer Interval 2	8	RW	Req	n/a	MR-IOV Cap / Watchdog
Watchdog 1 Expired	1	RW1C	Req	n/a	MR-IOV Cap / Watchdog
Rearm Watchdog 1 & 2	1	RW1C	Req	n/a	MR-IOV Cap / Watchdog
<b>Performance Monitoring</b>					
Statistics Interrupt Enable	1	RW	Opt	BF / Control	MR-IOV Cap / Control
Statistics Interrupt Status	1	RW1C	Opt	BF / Status	MR-IOV Cap / Status
# Statistics Blocks	8	RO	Req	BF / Statistics	MR-IOV Cap / Statistics
# Statistics Descriptors	8	RO	Req	BF / Statistics	MR-IOV Cap / Statistics
Statistics Block Start / Busy	up to 32	RW	Opt	BF / Statistics	MR-IOV Cap / Statistics
Statistics Block Offset / BIR	32	RO	Opt	BF / Statistics	MR-IOV Cap / Statistics
Statistics Descriptor Offset / BIR	32	RO	Opt	BF / Statistics	MR-IOV Cap / Statistics
<b>Link Control</b>					
MaxVH	8	RO	Req	Main BF / VH Counts	Port Table / Capability
NumVH	8	RW	Req	Main BF / VH Counts	Port Table / Control
Port Present	1	RO	Req	n/a	Port Table / Capability
Port Enable	1	RW	Req	n/a	Port Table / Control
Port Interrupt Enable	8	RW	Req	n/a	Port Table / Control
Port Interrupt Pending	8	RW1C	Req	n/a	Port Table / Status
MR-IOV Link Enable	1	RW	Req	n/a	Port Table / Control
Force Reset	1	RW	Req	n/a	VS Bridge Table / Hot Plug Signals
VS Interrupt Enable	1	RW	Req	n/a	VS Table / Control

Field Name	Width	Type	Opt	Device Usage	Switch Usage
VS Suppress Reset Propagation	1	RW	Req	n/a	VS Table / Control
PM_PME Triggers Beacon / WAKE#	1	RW	Opt	n/a	Port Table / Control
Send PME_Enter_L23 DLLP	1	RW	Opt	n/a	Port Table / Control
Non-PCIe Management Port	1	RO	Req	n/a	Port Table / Capability
MR Error Status	32	RW1C	Opt	Main BF / MR Log	Port Table / MR Log
MR Error Log	160 (5*32)	RO	Opt	Main BF / MR Log	Port Table / MR Log
Max Payload Size Supported	3	RO	Req	n/a	VS Bridge Table / Capability
Max Payload Size Offered	3	RW	Opt	n/a	VS Bridge Table / Control
Link Direction Status	2	RO	Req	n/a	Port Table / Status
Link Partner Detected	1	RO	Req	n/a	Port Table / Status
Link Direction Control + Backup Link Direction Control	2+2	RW	Req	n/a	Port Table / Control
Vendor Specific MR Init Information	3	RW	Req	n/a	Port Table / Control
Link Partner Vendor Specific MR Init Information	3	RO	Req	n/a	Port Table / Link Partner
Link Partner MaxVH	8	RO	Req	n/a	Port Table / Link Partner
Link Partner MaxVL	3	RO	Req	n/a	Port Table / Link Partner
Link Partner Trained in MR Mode	1	RO	Req	n/a	Port Table / Link Partner
Link Partner Protocol Version	3	RO	Req	n/a	Port Table / Link Partner
Link Partner was Authorized	1	RO	Req	n/a	Port Table / Link Partner
Link Partner Type	4	RO	Req	n/a	Port Table / Link Partner
<b>VF Migration / VF Mapping</b>					
VF Migration Supported	1	RO	Req	BF / Capability	n/a
VF Mapping Supported	1	RO	Req	BF/ Capability	n/a
VF Enable	1	RO	Req	Function Table / Status	n/a

Field Name	Width	Type	Opt	Device Usage	Switch Usage
VF Enable Changed	1	RW1C	Req	Function Table / Control	n/a
VF Enable Enabled	1	RW	Req	Function table / Control	n/a
VF Migration Capable	1	RW	Opt	Function Table / Control	n/a
VF Migration Enabled	1	RO	Opt	Function Table / Status	n/a
Max LVF #	16	RO	Opt	BF / VF Migration	n/a
Max MVF #	16	RO	Opt	BF / VF Migration	n/a
LVF Table Offset / BIR	32	RO	Opt	BF / VF Migration	n/a
Base LVF	16	RW	Opt	Function Table	n/a
InitialVFs	16	RW	Opt	Function Table / Control	n/a
TotalVFs	16	RW	Opt	Function Table / Control	n/a
NumVFs	16	RO	Opt	Function Table / Status	n/a
First VF Offset	16	RO	Opt	Function Table	n/a
VF Stride	16	RO	Opt	Function Table	n/a
VF Migration Status Interrupt Enable	1	RW	Opt	Function Table	n/a
VF Migration Status	1	RW1C	Opt	Function Table	n/a
PF Reset Indicated Interrupt Enable	1	RW1C	Opt	Function Table	n/a
PF Reset Indicated	1	RW1C	Opt	Function Table	n/a
<b>Congestion Management</b>					
VL Enable	8	RW	Req	Main BF / Control	Port Table / Control
VL Negotiation Pending	8	RO	Req	Main BF / Status	Port Table / Status
MaxVL	3	RW	Opt	BF / VL Arb	Port Table / VL Arb
BF VL	3	RW	Req	Main BF / Control	n/a
Default VL	3	RW	Opt	Main BF / Control	n/a
VL Arbitration Table Offset	30	RO	Opt	BF / VL Arb	Port Table / VL Arb
VL Arbitration Reference Clock	2	RO	Opt	BF / VL Arb	Port Table / VL Arb
VL Arbitration Capability	8	RO	Opt	BF / VL Arb	Port Table / VL Arb
Load VL Arbitration Table	1	RW	Opt	BF / VL Arb	Port Table / VL Arb

Field Name	Width	Type	Opt	Device Usage	Switch Usage
VL Arbitration Select	4	RW	Opt	BF / VL Arb	Port Table / VL Arb
VL Arbitration Status	1	RO	Opt	BF / VL Arb	Port Table / VL Arb
Max Time Slots	8	RO	Opt	BF / VL Arb	Port Table / VL Arb
VL Strict Priority Arbitration	8	RW	Opt	BF / VL Arb	Port Table / VL Arb
VC Capability Supported	1	RO	Req	Function Table / Capability	n/a
Num VC Resources Hardware Present	3	RO	Req	Function Table / Capability	n/a
Num MFVC Resources Hardware Present	3	RO	Req	Function Table / Capability	n/a
MFVC Capability Supported	1	RO	Req	Function Table / Capability	n/a
Extended VC Count	3	RW	Opt	Function Table / Control	VS Bridge Table / Control
Low Priority Extended VC Count	3	RW	Opt	Function Table / Control	VS Bridge Table / Control
VC Resource Enabled	8 (1*8)	RO	Opt	Main BF Function Table / VC State	VS Bridge Table / VC State
VC Resource VC Negation Pending	8 (1*8)	RO	Opt	Main BF Function Table / VS State	VS Bridge Table / VC State
VC ID	24 (3*8)	RO	Opt	Function Table / VC State	VS Bridge Table / VC State
TC to VC Map	64 (8*8)	RO	Opt	Function Table / VC State	VS Bridge Table / VC State
VC to VL Map	24 (3*8)	RW	Opt	Main BF Function Table / VC ID to VL Map	VS Bridge Table / VC ID to VL Map
VC Mapped	8 (1*8)	RW	Opt	Main BF Function Table / VC ID to VL Map	VS Bridge Table / VC ID to VL Map
VC Config Changed	1	RW1C	Opt	Main BF Function Table / Status	VS Bridge Table / Status
VC Config Interrupt Enable	1	RW	Opt	Main BF Function Table / Control	VS Bridge Table / Control
MFVC Resource Enabled	8 (1*8)	RO	Opt	Main BF Function Table / MFVC State	n/a
MFVC Resource VC Negation Pending	8 (1*8)	RO	Opt	Main BF Function Table / MFVC State	n/a
MFVC: VC ID	24 (3*8)	RO	Opt	Function Table / MFVC State	n/a
MFVC: TC to VC Map	64 (8*8)	RO	Opt	Function Table / MFCV	n/a



Field Name	Width	Type	Opt	Device Usage	Switch Usage
				State	
MFVC Config Changed	1	RW1C	Opt	Main BF Function Table / Status	n/a
MFVC Config Interrupt Enable	1	RW	Opt	Main BF Function Table / Control	n/a
<b>Global Key Management</b>					
Global Key Value	12	RW	Req	Main BF / Function Table	VS Table / Control
Device Global Key Check Enable	1	RW	Req	Main BF / Control	n/a
VS Global Key Check Enable	3	RW	Req	n/a	VS Table / Control
<b>Management Authorization</b>					
Management VS	8	RW	Req	n/a	MR-IOV Cap / Control
Authorized VS Bitmap Offset	12	RO	Req	n/a	MR-IOV Cap / Control
Authorized VS Bitmap + Backup Authorized Bitmap	2 * NumVS	RW	Req	n/a	MR-IOV Cap / Control
<b>Switch Mapping</b>					
Bridge Hardware Present	1	RO	Req	n/a	VS Bridge Table / Capability
Bridge Enable	1	RW	Req	n/a	VS Bridge Table / Control
Bridge Port	8	RW	Req	n/a	VS Bridge Table / Control
Bridge Port VH	12	RW	Req	n/a	VS Bridge Table / Control
Port Mapped to Bridge	1	RW	Req	n/a	VS Bridge Table / Control
VS Present	1	RO	Req	n/a	VS Table / Capability
VS Enable	1	RW	Req	n/a	VS Table / Control
VS Interrupt Vector Num	11	RO	Req	n/a	VS Table / Capability
<b>Hot Plug Signals Interface</b>					
Hot Plug Hardware Present	1	RO	Req	n/a	VS Bridge Table / Capability
Bridge Controls Physical Link	1	RW	Req	n/a	VS Bridge Table / Hot Plug Signals
Slot Implemented	1	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals

Field Name	Width	Type	Opt	Device Usage	Switch Usage
Virtual Hot Plug Interrupt Enable	1	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals
PME Turn Off State	1	RO	Req	n/a	VS Bridge Table / Status
Physical Slot Number	13	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals
Slot Power Limit Scale	2	RO	Opt	n/a	VS Bridge Table / Hot Plug Signals
Slot Power Limit Value	8	RO	Opt	n/a	VS Bridge Table / Hot Plug Signals
Hot Plug Capable	1	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals
Hot Plug Surprise	1	RW	Opt	n/a	VS Bridge Table / hot Plug Signals
Attention Indicator State	2	RO	Opt	n/a	VS Bridge Table / Hot Plug Signals
Attention Indicator Changed	1	RW1C	Opt	n/a	VS Bridge Table / Status
Push Attention Button	1	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals
Power Indicator State	2	RO	Opt	n/a	VS Bridge Table / Hot Plug Signals
Power Indicator Changed	1	RW1C	Opt	n/a	VS Bridge Table / Status
Power Controller State	1	RO	Opt	n/a	VS Bridge Table / Hot Plug Signals
Power Controller Changed	1	RW1C	Opt	n/a	VS Bridge Table / Status
Power Controller Present	1	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals
Signal Power Fault	1	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals
Presence Detect State	1	RW	Opt	n/a	VS Bridge Table / Hot Plug Signals

## 4.2. Device Configuration Space

For managing MR-IOV Devices, MR-PCIM uses MR-IOV Capabilities located in the Base Functions. Base Functions are Type 0 Headers located in VH0 of the Device that contain the MR-IOV Capability.

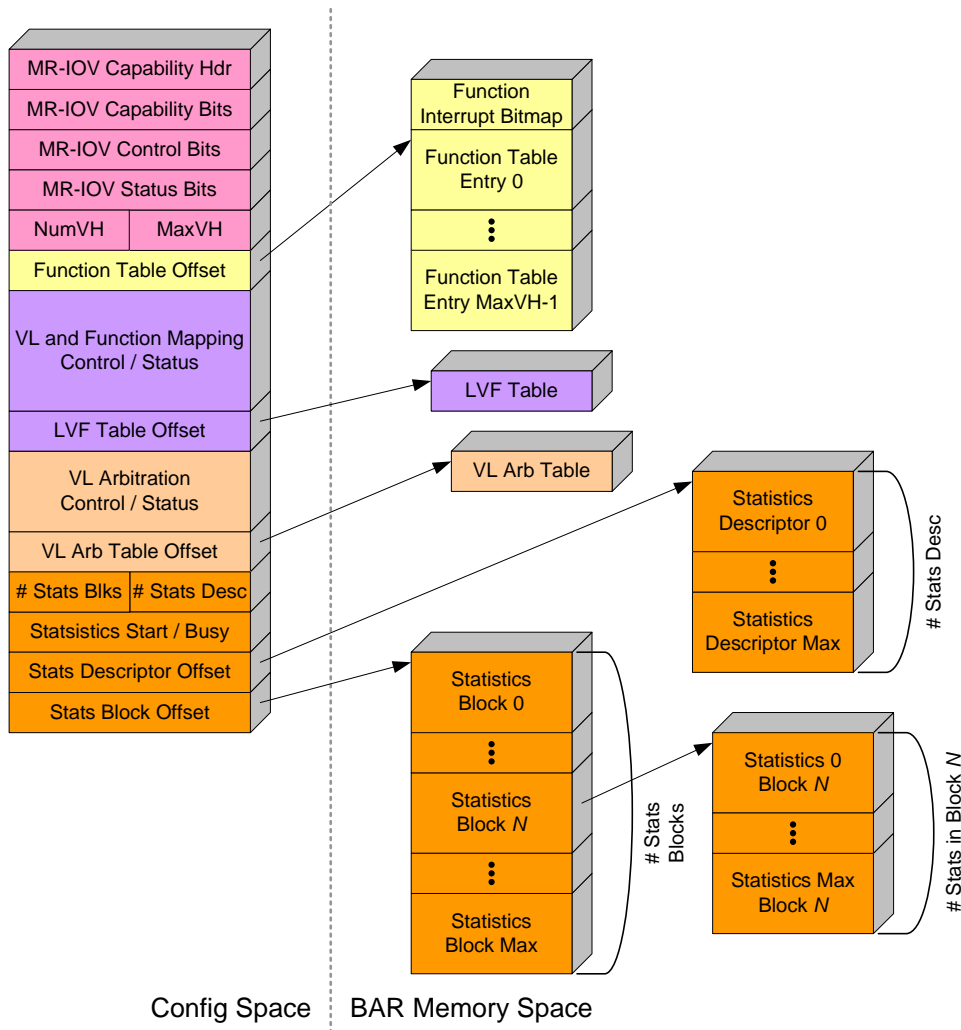
Configuration controls are associated with:

- ❑ Entire component
- ❑ Each Base Function
- ❑ Each VH supported by the component
- ❑ Each Function or PF supported in some non-management VH by the component

There are up to five tables provided by each Base Function. These tables are located using the MR-IOV Capability block located in the Type 0 Configuration Space of the associated BF. An overview of the tables is shown in [Figure 4-1](#).

Deleted: Figure 4-1

- ❑ The MR-IOV Capability contains information concerning the BF.
- ❑ The VH Table contains information concerning each VH supported by the Device. There is exactly one VH Table per Device and it is associated with an arbitrary BF.
- ❑ The optional VL Arbitration Table contains information describing how VL Arbitration is performed by the Device. If present, this table is associated with the BF that contains the VH Table.
- ❑ The optional LVF Table is used to control VF Mapping and VF Migration. This table is absent if neither VF Mapping nor VF Migration are supported by the BF.
- ❑ The Function Table contains one entry for each Function or PF in each non-management VH.



**Figure 4-1: MR Device Configuration Space**

#### 4.2.1. Device MR-IOV Extended Capability

MR-IOV Base Functions contain a new PCIe Extended Capability. This capability is used by MR-PCIM to determine if a Device is MR-IOV capable and to manage the MR-IOV features of the Function.

The MR-IOV capability is located in the PCIe Extended Configuration Space. [Figure 4-2](#) shows the Device MR-IOV Extended Capability structure:

Deleted: Figure 4-2

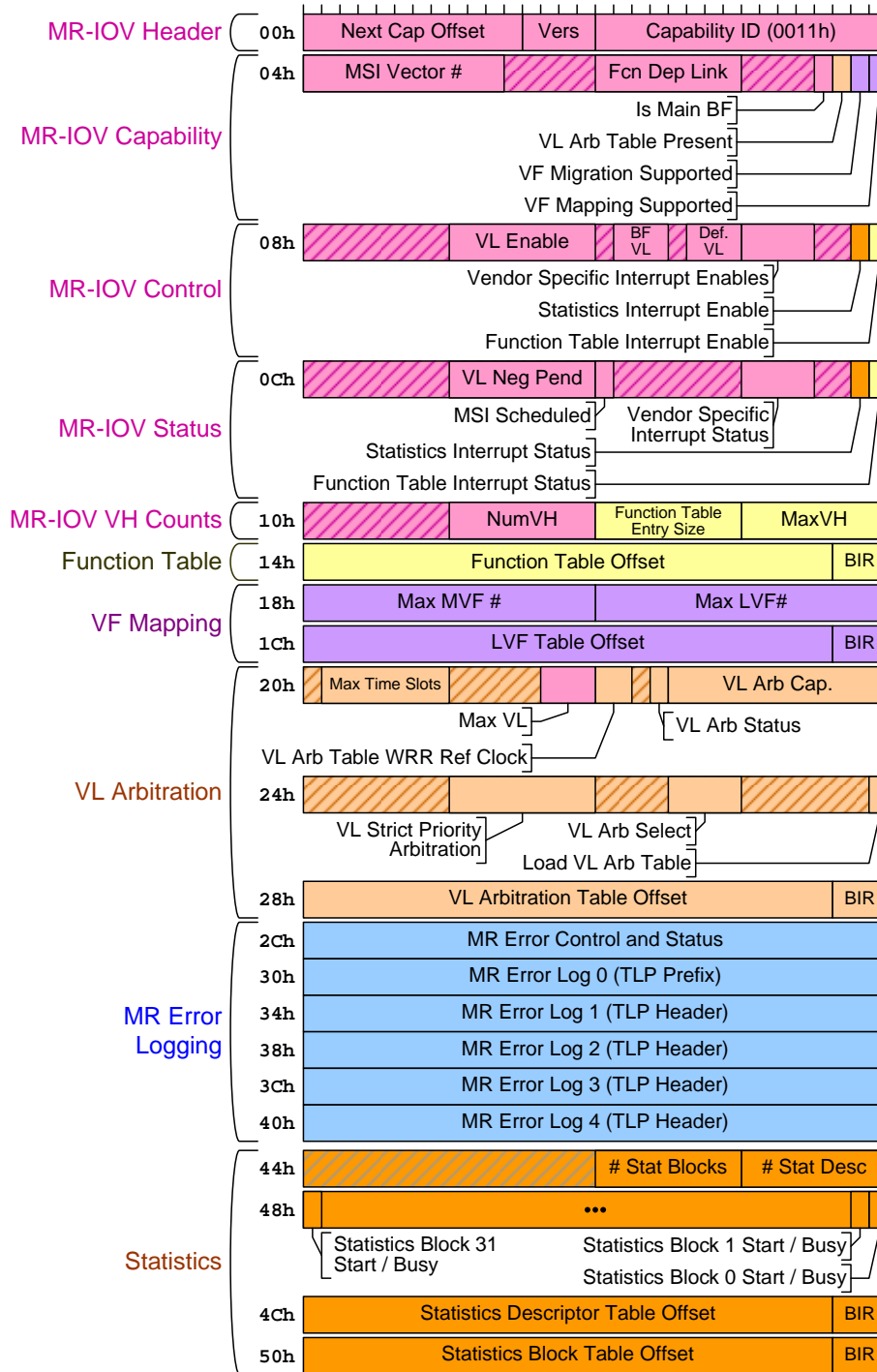


Figure 4-2: Device MR-IOV Capability

4.2.1.1. MR-IOV Extended Capability Header (00h)

Table 4-2 defines the Switch MR-IOV Extended Capability header. The Capability ID for the Switch MR-IOV Extended Capability is 0011h.

Deleted: Table 4-2

Table 4-2: Device MR-IOV Extended Capability Header

Bit Location	Register Description	Attributes
15:0	<b>PCI Express Extended Capability ID</b> – This field is a PCI-SIG defined ID number that indicates the nature and format of the Extended Capability.  The Extended Capability ID for the MR-IOV Extended Capability is 0011h.	RO
19:16	<b>Capability Version</b> – This field is a PCI-SIG defined version number that indicates the version of the Capability structure present.  Must be 1h for this version of the specification.	RO
31:20	<b>Next Capability Offset</b> – This field contains the offset to the next PCI Express Capability structure or 000h if no other items exist in the linked list of Capabilities.  This offset is relative to the beginning of PCI compatible Configuration Space and thus must always be either 000h (for terminating list of Capabilities) or greater than 0FFh.	RO

#### 4.2.1.2. MR-IOV Capabilities (04h)

Table 4-3: MR-IOV Capabilities

Bit Location	Register Description	Attributes
0	<b>VF Mapping Supported</b> – If Set, this BF supports mapping of Mission Virtual Functions (MVs) into VFs. If Clear, any mapping is Vendor Specific and is not controlled by MR-PCIM.	RO
1	<b>VF Migration Supported</b> – If Set, this BF supports dynamic mapping of Mission Virtual Functions into and out of VFs. If Clear, VF Migration support is not available.  If VF Migration is Set, VF Mapping must also be Set.  VF Migration can not occur unless enabled by software. Migration is enabled on a per-PF basis. See Section xxx for details.	RO
2	<b>VL Arbitration Table Present</b> – If Set, the VL Arbitration Table is present. If Clear, the VL Arbitration Table is not present and the corresponding fields in this BF are Read Only Zero.  This field is only meaningful in the Main BF of the Device. In all other BFs, this field is Read Only Zero.	RO
3	<b>Is Main BF</b> – If Set, this BF is the “Main” BF of the Device. Certain fields are only meaningful in the Main BF.	RO
7:4	<b>Reserved</b>	RO
15:8	<b>Function Dependency Link</b> –The programming model for a Device may have Vendor Specific dependencies between sets of Functions. The Function Dependency Link field is used to inform MR-PCIM about these dependencies.  This field describes dependencies between BFs. PF and VF dependencies are the same as the dependencies of their associated BFs.  If a BF is independent from other BFs of a Device, this field shall contain the Function Number of the BF.  If a BF is dependent on other BFs of a Device, this field shall contain the Function Number of the next BF in the same Function Dependency List. The last BF in a Function Dependency List shall contain the Function Number of the first BF in the Function Dependency List.  For BFs in a Function Dependency List, MR-PCIM must allocate MVs consistently. For these BFs, every VH should contain the exact same mapping of MVs to VFs.	RO
20:16	<b>Reserved</b>	RO
31:21	<b>MSI Vector Number</b> – This field indicates the MSI Vector number used to signal MR-IOV events within this BF. This value may change based on whether MSI or MSI-X is enabled.	RO

#### 4.2.1.3. MR-IOV Control (08h)

Table 4-4: Device MR-IOV Control

Bit Location	Register Description	Attributes
0	<b>Function Table Interrupt Enable</b> – If Set, when any bit in the Function Interrupt Bitmap is Set, an interrupt is requested. If Clear, transitions of the Function Interrupt Bitmap do not request an interrupt. Default is 0b.	RW
3:1	<b>Reserved</b>	RO
7:4	<b>Vendor Specific Interrupt Enables</b> – If Set, various Vendor Specific events can cause an interrupt. If Clear, these Vendor Specific Events can not cause an interrupt.  Vendor Specific Interrupt support is optional. These bits are Read Only Zero for any unimplemented Vendor Specific Interrupts. Default is 0b.	RW
10:9	<b>Default VL</b> – VL Number used for TLPs originated in VH0 and not associated with a BF, PF or VF (i.e. a “Plain Old Function”). Such TLPs must use TC 0 / VC 0.  This field is only meaningful in the Main BF of the Device. In all other BFs, it is Read Only Zero.  If all functions in VH0 are BFs, PFs or VFs, this field may be Read Only Zero.	RW
11	<b>Reserved</b>	RO
14:12	<b>BF VL</b> – VL Number used for TLPs originating in this BF. Such TLPs must use TC 0 / VC 0.  This field is only meaningful in the Main BF of the Device. In all other BFs, it is Read Only Zero.	RW
15	<b>Reserved</b>	RO
23:16	<b>VL Enable</b> – This bit, when Set, enables a Virtual Link – This bit, when Set, enables a Virtual Link (see note 1 for exceptions). The Virtual Link is disabled when this bit is cleared.  Software must use the VL Negotiation Pending bit to check whether the VL negotiation is complete.  Default value of this bit is 1b for VL0 and is 0b for other VLs.  The number of bits implemented in this field is determined by the MaxVL value. VL Enable bits for VLs greater than MaxVL are Read Only Zero.  This field is only meaningful in the Main BF of the Device. In all other BFs, it is Read Only Zero.  Notes: 1. This bit is hardwired to 1b for the VL0, i.e., writing to this bit has no effect for VL0. 2. To enable a Virtual Link, the VL Enable bits for that Virtual Link	RW



Bit Location	Register Description	Attributes
	must be Set in both components on a Link. 3. To disable a Virtual Link, the VL Enable bits for that Virtual Link must be cleared in both components on a Link. 4. Software must ensure that no traffic is using a Virtual Link at the time it is disabled. 5. Software must fully disable a Virtual Link in both components on a Link before re-enabling the Virtual Link.	
31:24	<b>Reserved</b>	RO

#### 4.2.1.4. MR-IOV Status (0Ch)

Table 4-5: Device MR-IOV Status

Bit Location	Register Description	Attributes
0	<b>Function Table Interrupt Status</b> – Set if any bit in the Function Interrupt Bitmap is Set.	RO
3:1	<b>Reserved</b>	RO
7:4	<p><b>Vendor Specific Interrupt Status</b> – Set if the corresponding Vendor Specific Interrupt condition has occurred.</p> <p>Vendor Specific Interrupt support is optional. These bits are Read Only Zero for any unimplemented Vendor Specific Interrupts.</p> <p>Vendor Specific Interrupt support allows software to read a single register in response to an interrupt and quickly determine the reason for the interrupt.</p> <p>It is Vendor Specific how a Vendor Specific Interrupt is cleared. Clearing occurs either due to writing 1b values to this field or by clearing the interrupt using other, Vendor Specific fields.</p>	RO / RW1C
14:8	<b>Reserved</b>	RO
15	<b>MSI Scheduled</b> – Set when an MSI has been requested. If Set, subsequent MSIs are suppressed. If Clear, any enabled interrupt will cause an MSI to be scheduled (and this bit to be Set). Default is 0b.	RW1C
23:16	<p><b>VL Negotiation Pending</b> – These bits indicate whether Flow Control negotiation for a VL is in the pending state.</p> <p>The value of this bit is defined only when the Link is in the DL_Active state and the Virtual Link is enabled (its VL Enable bit is Set).</p> <p>When this bit is Set by hardware, it indicates that the VL resource has not completed the process of negotiation. This bit is cleared by hardware after the VL negotiation is complete (on exit from the MR_FC_INIT2 state on the VL).</p> <p>This field is only meaningful in the Main BF of the Device. In all other BFs, this field is Read Only Zero,</p>	RO
31:24	<b>Reserved</b>	RO

#### 4.2.1.5. MR-IOV VH Counts (10h)

Table 4-6: Device MR-IOV VH Counts

Bit Location	Register Description	Attributes
7:0	<b>MaxVH</b> – Maximum number of VHs supported on this Device. VHs are numbered [0..MaxVH-1].  This field is only meaningful in the Main BF of the Device. In all other BFs, this field is Read Only Zero,	RO
15:8	<b>Function Table Entry Size</b> – Returns the size, in DWORDs of each Function Table Entry. For this version of the specification, this value is at least 16.	RO
23:16	<b>NumVH</b> – Indicates the number of VHs enabled by the upstream component. This value must be less than or equal to MaxVH.. The default value of this field is 0 indicating one VH is enabled.  This field is only meaningful in the Main BF of the Device. In all other BFs, this field is Read Only Zero,	RW
31:24	<b>ReservedZ</b>	RO

#### 4.2.1.6. Function Table Offset (14h)

Table 4-7: Device Function Table Offset

Bit Location	Register Description	Attributes																					
2:0	<b>Function Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's Function Table into Memory Space.  <b>BIR Value Base Address register</b>  <table> <tr><td>0</td><td>BAR0</td><td>10h</td></tr> <tr><td>1</td><td>BAR1</td><td>14h</td></tr> <tr><td>2</td><td>BAR2</td><td>18h</td></tr> <tr><td>3</td><td>BAR3</td><td>1Ch</td></tr> <tr><td>4</td><td>BAR4</td><td>20h</td></tr> <tr><td>5</td><td>BAR5</td><td>24h</td></tr> <tr><td>6..7</td><td>Reserved</td><td></td></tr> </table> For a 64-bit Base Address register, the BIR indicates the lower DWORD.	0	BAR0	10h	1	BAR1	14h	2	BAR2	18h	3	BAR3	1Ch	4	BAR4	20h	5	BAR5	24h	6..7	Reserved		RO
0	BAR0	10h																					
1	BAR1	14h																					
2	BAR2	18h																					
3	BAR3	1Ch																					
4	BAR4	20h																					
5	BAR5	24h																					
6..7	Reserved																						
31:3	<b>Function Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the Function Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.	RO																					

The total size of the table (in bytes) is:

$$((\text{MaxVH} + 1) * \text{Function\_Table\_Entry\_Size} * 4) + 32$$

This includes both the Function Table and the Function Interrupt Bitmap.

The Function Interrupt Bitmap immediately precedes the Function Table. The size of the Function Table Interrupt Bitmap is 32 bytes (supporting a maximum of 256 VHs). Bit 0 of the first DWORD corresponds to VH 0; bit 1 corresponds to VH 1, ... Any unused bits are Read Only Zero.

#### 4.2.1.7. MVF and LVF Sizes (18h)

If VF Mapping is supported, this register indicates the *MVF Index* values that may be assigned to a VF. Values in the range [1..Max MVF] may be written to the LVF Table mapping field.

**Table 4-8: VF MVF Region**

Bit Location	Register Description	Attributes
15:0	<p><b>Max LVF</b> – Indicates size of the LVF Table. The LVF Table contains MaxLVF+1 entries numbered 0 to Max LVF.</p> <p>If VF Mapping is not supported, this field is Zero. All BFs from the same Function Dependency Group must have the same Max LVF value.</p>	RO
31:15	<p><b>Max MVF</b> – Indicates the highest MVF Index that can be assigned to a VF. Zero if VF Mapping Supported is Clear. MVF numbers are in the range [1..MaxMVX #].</p> <p>If VF Mapping is not supported, this field is Zero. All BFs from the same Function Dependency Group must have the same Max MVF value.</p>	RO

#### 4.2.1.8. LVF Table Offset (1Ch)

Table 4-9: LVF Table Offset

Bit Location	Register Description	Attributes																					
2:0	<p><b>LVF Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's LVF Table into Memory Space.</p> <p><b>BIR Value Base Address register</b></p> <table> <tr><td>0</td><td>BAR0</td><td>10h</td></tr> <tr><td>1</td><td>BAR1</td><td>14h</td></tr> <tr><td>2</td><td>BAR2</td><td>18h</td></tr> <tr><td>3</td><td>BAR3</td><td>1Ch</td></tr> <tr><td>4</td><td>BAR4</td><td>20h</td></tr> <tr><td>5</td><td>BAR5</td><td>24h</td></tr> <tr><td>6..7</td><td>Reserved</td><td></td></tr> </table> <p>For a 64-bit Base Address register, the BIR indicates the lower DWORD.</p>	0	BAR0	10h	1	BAR1	14h	2	BAR2	18h	3	BAR3	1Ch	4	BAR4	20h	5	BAR5	24h	6..7	Reserved		RO
0	BAR0	10h																					
1	BAR1	14h																					
2	BAR2	18h																					
3	BAR3	1Ch																					
4	BAR4	20h																					
5	BAR5	24h																					
6..7	Reserved																						
31:3	<p><b>LVF Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the VH Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.</p>	RO																					

The total size of the LVF Table (in bytes) is:                      Max LVF \* 4

#### 4.2.1.9. VL Arbitration Capability and Status (20h)

Table 4-10: Device VL Arbitration Capability and Status

Bit Location	Register Description	Attributes
11:0	<p><b>VL Arbitration Capability</b> – Indicates the types of VL Arbitration supported by the Port. This field is valid for all Functions that report a Low Priority Extended VC Count field greater than 0. For all other Functions, this field must be hardwired to 00h.</p> <p>Each bit location within this field corresponds to a VC Arbitration Capability defined below. When more than 1 bit in this field is Set, it indicates that the Port can be configured to provide different VC arbitration services. Defined bit positions are:</p> <p>Bit 0      Hardware fixed arbitration scheme, e.g., Round Robin</p> <p>Bit 1      Weighted Round Robin (WRR) arbitration with 32 phases</p> <p>Bit 2      WRR arbitration with 64 phases</p> <p>Bit 3      WRR arbitration with 128 phases</p> <p>Bit 4      Time-based WRR with 128 phases</p> <p>Bit 5      WRR Arbitration with 256 phases</p> <p>Bits 6-7   Reserved</p> <p>Bit 10-8   Vendor Defined VL Arbitration Scheme</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	RO
12	<p><b>VL Arb Status</b> – This bit indicates the coherency status of the VL Arbitration Table. This bit is valid only when the VL Arbitration Table is used.</p> <p>This bit is Set by hardware when any entry of the VL Arbitration Table is written to by software. This bit is cleared by hardware when hardware finishes loading values stored in the VL Arbitration Table after software sets the Load VL Arbitration Table bit.</p> <p>Default value of this bit is 0b.</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	RO
13	<b>Reserved</b>	RO
15:14	<p><b>Reference Clock</b> – Indicates the reference clock for Virtual Links that support time-based WRR VL Arbitration. This field is valid only if time-based WRR is supported.</p> <p>Defined encodings are:</p> <p>00b      100 ns reference clock</p> <p>01b – 11b   Reserved</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	RO
18:16	<p><b>MaxVL</b> – Indicates the number of VLs supported. The Device supports VL<sub>0</sub> through VL<sub>MaxVL</sub> inclusive. This field is only meaningful in the Main BF of the Device. In all other BFs, it is Read Only Zero.</p>	RO

Bit Location	Register Description	Attributes
23:19	<b>Reserved</b>	RO
30:24	<b>Maximum Time Slots</b> – Indicates the maximum number of time slots (minus one) that are supported when configured for time-based WRR VL Arbitration. For example, a value 000 0000b in this field indicates the supported maximum number of time slots is 1 and a value of 111 1111b indicates the supported maximum number of time slot is 128.  This field is valid only when the VL Arbitration Capability field indicates that time-based WRR VL Arbitration is supported.  This field is Read Only Zero if VL Arbitration Present is Clear.	RO
31	<b>Reserved</b>	RO

#### 4.2.1.10. VL Arbitration Control (24h)

Table 4-11: Device VL Arbitration Control

Bit Location	Register Description	Attributes
0	<p><b>Load VL Arbitration Table</b> – When Set, this bit updates the VL Arbitration logic from the VL Arbitration Table. This bit is valid only when the VL Arbitration Table is used by the selected VL Arbitration scheme (that is indicated by a Set bit in the VL Arbitration Capability field selected by VL Arbitration Select).</p> <p>Software sets this bit to signal hardware to update VL Arbitration logic with new values stored in VL Arbitration Table; clearing this bit has no effect. Software uses the VL Arbitration Table Status bit to confirm whether the new values of VL Arbitration Table are completely latched by the arbitration logic.</p> <p>This bit always returns 0b when read.</p> <p>Default value of this bit is 0b.</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	RW
7:4	<b>Reserved</b>	RO
11:8	<p><b>VL Arbitration Select</b> – This field configures the Port to provide a particular VL Arbitration service.</p> <p>The permissible value of this field is a number corresponding to one of the asserted bits in the VL Arbitration Capability field.</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	RW
15:12	<b>Reserved</b>	RO
23:16	<p><b>VL Strict Priority Arbitration</b> – This field contains one bit per VL. Bit 0 corresponds to VL0. Bit 7 corresponds to VL7.</p> <p>When a bit is Set, the corresponding VL is configured to arbitrate as Strict Priority based on VL number. When a bit is Clear, the corresponding VL is configured to arbitrate as normal priority (using the scheme selected by VL Arbitration Select).</p> <p>Among the VLs configured for strict priority, priority is based on increasing VL number. VL0 is the lowest strict priority, VL7 is the highest.</p> <p>Strict Priority VLs have priority over normal priority VLs.</p> <p>Behavior is Undefined if a VL configured for Strict Priority is also included in the VL Arbitration Table.</p> <p>If a VL is Disabled, the value of the corresponding bit in this field is ignored.</p> <p>Default value of this field is 0000 0000b.</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	RW
31:24	<b>Reserved</b>	RO



4.2.1.11. VL Arbitration Table Offset (28h)

Table 4-12: Device VL Arbitration Table Offset

Bit Location	Register Description	Attributes																					
2:0	<p><b>VL Arbitration Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's VL Arbitration Table into Memory Space.</p> <p><b>BIR Value Base Address register</b></p> <table> <tr><td>0</td><td>BAR0</td><td>10h</td></tr> <tr><td>1</td><td>BAR1</td><td>14h</td></tr> <tr><td>2</td><td>BAR2</td><td>18h</td></tr> <tr><td>3</td><td>BAR3</td><td>1Ch</td></tr> <tr><td>4</td><td>BAR4</td><td>20h</td></tr> <tr><td>5</td><td>BAR5</td><td>24h</td></tr> <tr><td>6..7</td><td>Reserved</td><td></td></tr> </table> <p>For a 64-bit Base Address register, the BIR indicates the lower DWORD.</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	0	BAR0	10h	1	BAR1	14h	2	BAR2	18h	3	BAR3	1Ch	4	BAR4	20h	5	BAR5	24h	6..7	Reserved		RO
0	BAR0	10h																					
1	BAR1	14h																					
2	BAR2	18h																					
3	BAR3	1Ch																					
4	BAR4	20h																					
5	BAR5	24h																					
6..7	Reserved																						
31:3	<p><b>VL Arbitration Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the VL Arbitration Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.</p> <p>This field is Read Only Zero if VL Arbitration Present is Clear.</p>	RO																					

#### 4.2.1.12. MR Error Status (2Ch)

Table 4-13: Device MR Error Status

Bit Location	Register Description	Attributes
3:0	MR First Error Pointer	RO
4	MR Uncorrectable TLP Error Status	RW1C
5	MR Global Key Error Status	RW1C
14:6	Reserved	RsvdZ
15	MR DLLP Error Status – DLLP Errors are not logged.	RW1C

#### 4.2.1.13. MR Error Control (2Eh)

Table 4-14: Device MR Error Control

Bit Location	Register Description	Attributes
3:0	Reserved	RO
4	MR Uncorrectable TLP Error Mask	RW
5	MR Global Key Error Mask	RW
14:6	Reserved	RO
15	MR DLLP Error Mask	RW

#### 4.2.1.14. MR Error Log (30h to 40h)

These fields contain the TLP Prefix and TLP Header corresponding to the error described by the First Error Pointer in the MR Error Status register.

The value of these fields is undefined if the First Error Pointer is zero or points to a bit number that is not Set.

Headers are not logged and the First Error Pointer is not updated for DLLP Errors.

#### 4.2.1.15. Statistics Capability and Control (44h to 50h)

Device and Switch Statistics related fields are described in Section 4.3.7.

### 4.2.2. Device VL Arbitration Table

Switch and Device VL Arbitration tables are identical. See Section 4.3.7 for details.

### 4.2.3. LVF Table

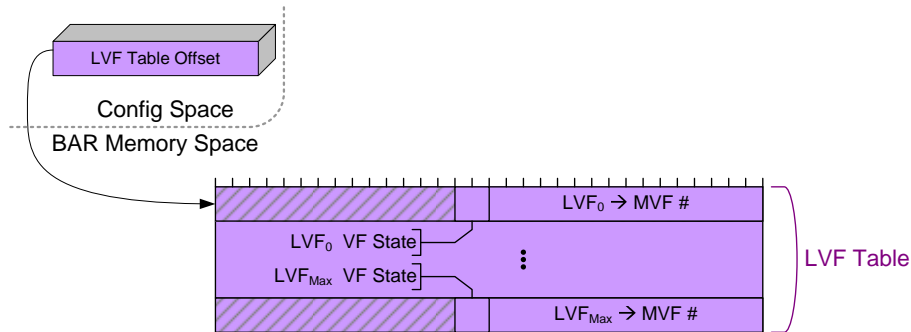


Figure 4-3: LVF Table

The LVF Table is present if VF Mapping Supported is Set.

LVF Table Entries are one DWORD.

#### 4.2.3.1. LVF Table Entry

Table 4-15: LVF Table Entry

Bit Location	Register Description	Attributes
15:0	<b>MVF # for this LVF</b> – MVF mapped to this LVF. Only enough bits are implemented to express values in the range [0..VF MVF Region High]. Unimplemented high order bits are read only zero.  The value 0 indicates this LVF is not mapped to any MVF.  Default value is Vendor Specific.	RW
17:16	<b>VF State</b> – VF Migration State for this LVF. Values are: 00 Inactive.Unavailable 01 Inactive.MigrateIn 11 Active.Available 10 Active.MigrateOut  MR-PCIM changes VF Migration state by writing this value.  If VF Migration is not supported, this field is Read Only Zero.  Even when VF Migration is Disabled, MR-PCIM must ensure that this field is meaningful (i.e. VFs below InitialVFs should be Active.Available).	RW
31:18	<b>Reserved</b>	RO

#### 4.2.4. Function Table

The Function table contains one entry for each VH. The first entry represents VH0; the second entry represents VH1, etc.

Every non-zero VH contains exactly one PF associated with this BF. VH0 can optionally have a single PF.

Certain Function Table Entry fields are only implemented in the Main BF of the Device (for example, the Global Key fields and MFVC related fields).

The first Function Table Entry is always present. If VH0 does not contain a PF, most of the fields are Read Only Zero but a few fields remain meaningful (e.g. Global Key).

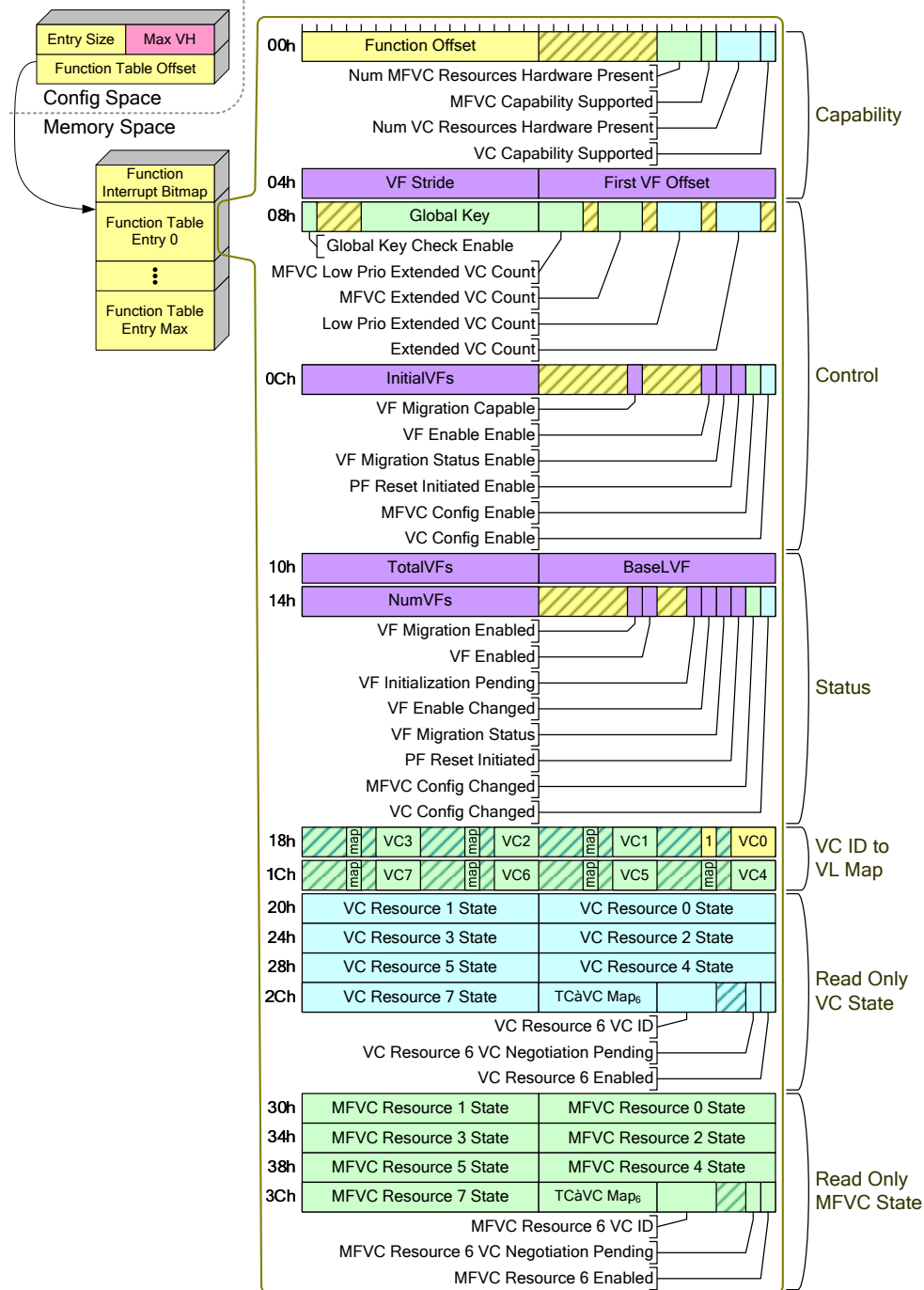


Figure 4-4: Device Function Table

4.2.4.1. Function Capability (00h and 04h)

Table 4-16: Function Capability 1 (00h)

Bit Location	Register Description	Attributes
0	<b>VC Capability Supported</b> –If Set, the Function contains a VC Capability. If Clear, the Function does not contain a VC Capability.	RO
3:1	<b>Num VC Resources Hardware Present</b> –Number of VCs supported in hardware in each VH's VC. Maximum value of this field is MaxVL. If VC Capability Supported is Clear, this field is hardwired 000b.	RO
4	<b>MFVC Capability Supported</b> – If Set, in every VH, the PF associated with this BF contains an MFVC Capability. If Clear, the PF associated with this BF does not contain an MFVC Capability.  This bit may only be Set in the BF that manages Function 0 within the VH (i.e. Function Offset is 0000h).	RO
7:5	<b>Num MFVC Resources Hardware Present</b> –Number of VCs supported in hardware in each VH's MFVC. Maximum value of this field is MaxVL. If MFVC Capability Supported is Clear, this field is hardwired 000b.	RO
15:8	<b>Reserved</b>	RO
31:16	<b>Function Offset</b> – Offset that describes the PF's RID within the associated VH. If the Captured Bus Number is NNh, the PF RID is Function Offset plus NN00h.  For the first Function Table Entry (i.e. VH0), Function Offset must be in the range [0000h..01FFh] or FFFFh. The values 0000h to 01FFh allow the VH0 PF to be located anywhere on either the Captured Bus Number or the Captured Bus Number + 1. The special FFFFh value indicates that this BF does not have a PF in VH0.  For the remaining Function Table Entries, Function Offset must be in the range [0000h..00FFh]. these values allow the associated PF to be located anywhere on the captured Bus Number.  With the exception of the first Function Table Entry (i.e. VH0), all Function Offset values for a given BF must have the same value.	RO

**Comment [sdg6]:** 0100h for VH0?  
What do we support for MFVC in the case where PFs are on two bus #s?

Table 4-17: Function Capability 2 (04h)

Bit Location	Register Description	Attributes
15:0	<b>First VF</b> – Contains the First VF of this PF's SR-IOV Capability. Zero if this Function is not a PF (i.e. it has no SR-IOV Capability).	RO
31:16	<b>VF Stride</b> – Contains the VF Stride of this PF's SR-IOV Capability. Zero if this Function is not a PF (i.e. it has no SR-IOV Capability).	RO

4.2.4.2. Function Control (08h to 10h)

Table 4-18: Function Control 1 (08h)

Bit Location	Register Description	Attributes
0	<b>Reserved</b>	RO
3:1	<p><b>VC Extended VC Count</b> – The value presented to software in the VH in the Extended VC Count field of the MFVC Capability. Valid values are [0..Num VC Hardware Resources Present].</p> <p>MR-PCIM may set this value to restrict the number of VCs offered to a VH.</p> <p>Default value is 0h. If VC Capability Supported is Clear, this field is hardwired 0h.</p>	RW
4	<b>Reserved</b>	RO
7:5	<p><b>VC Low Priority Extended VC Count</b> – The value presented to software in the VH in the Low Priority Extended VC Count field of the VC Capability. Valid values are [0..VC Extended VC Count].</p> <p>The value of this field does not affect arbitration in any manner. This field is allows MR-PCIM to indicate to software in the VH which VCs it should think are strict priority arbitration.</p> <p>Default value is 0h. If VC Capability Supported is Clear, this field is hardwired 0h.</p>	RW
8	<b>Reserved</b>	RO
11:9	<p><b>MFVC Extended VC Count</b> – The value presented to software in the VH in the Extended VC Count field of the MFVC Capability. Valid values are [0..Num MFVC Hardware Resources Present].</p> <p>MR-PCIM may set this value to restrict the number of VCs offered to a VH.</p> <p>Default value is 000b. If MFVC Capability Supported is Clear, this field is hardwired 000b.</p>	RW
12	<b>Reserved</b>	RO
15:13	<p><b>MFVC Low Priority Extended VC Count</b> – The value presented to software in the VH in the Low Priority Extended VC Count field of the MFVC Capability. Valid values are [0..MFVC Extended VC Count].</p> <p>The value of this field does not affect arbitration in any manner. This field is allows MR-PCIM to indicate to software in the VH which VCs it should think are strict priority arbitration.</p> <p>Default value is 000b. If MFVC Capability Supported is Clear, this field is hardwired 000b.</p>	RW
27:16	<p><b>Global Key</b> – TLPs received are checked against this value. TLPs sent contain this value. If this field contains 000h, the "wild card" value, checking is disabled. If a TLP contains 000h, the TLP is a wild card TLP and checking always passes.</p> <p>Default value of this field is 000h.</p>	RW

Bit Location	Register Description	Attributes
	This field is only implemented in one BF of the Device. For that BF, all Function Table entries are implemented (even if VH0 the BF that manages Function 0 within the non-zero VHs (i.e. Function Offset is 0000h).	
30:28	<b>Reserved</b>	RO
31	<b>Global Key Check</b> Enable – If Set, Global Key checking is performed. If Clear, Global Key mismatches are ignored	RW



Table 4-19: Function Control 2 (0Ch)

Bit Location	Register Description	Attributes
0	<b>VC Config Interrupt Enable</b> – If Set, setting the VC Config Changed bit for this Function Table Entry sets the corresponding Function Interrupt Bitmap. If Clear, the Function Interrupt Bitmap associated with this Function Table Entry is not affected by VC Config Changed.  Default is 0b. This field is Read Only Zero if VC Capability Supported is Clear.	RW
1	<b>MFVC Config Changed Enable</b> – If Set, setting the MFVC Config Changed bit for this Function Table Entry sets the corresponding Function Interrupt Bitmap. If Clear, the Function Interrupt Bitmap associated with this Function Table Entry is not affected by MFVC Config Changed.  Default is 0b. This field is Read Only Zero if MFVC Capability Supported is Clear.	RW
2	<b>PF Reset Initiated Enable</b> – If Set, setting the PF Reset Initiated bit for this Function Table Entry sets the corresponding Function Interrupt Bitmap. If Clear, the Function Interrupt Bitmap associated with this Function Table Entry is not affected by PF Reset Initiated.  Default is 0b. This field is Read Only Zero if VF Migration Supported is Clear	RW
3	<b>VF Migration Status Enable</b> – If Set, setting the VF Migration Status bit for this Function Table Entry sets the corresponding Function Interrupt Bitmap. If Clear, the Function Interrupt Bitmap associated with this Function Table Entry is not affected by VF Migration State.  Default is 0b. This field is Read Only Zero if VF Migration Supported is Clear.	RW
4	<b>VF Enable Enable</b> – If Set, setting the VF Enable changed bit for this Function Table Entry sets the corresponding Function Interrupt Bitmap. If Clear, the Function Interrupt Bitmap associated with this Function Table Entry is not affected by VF Enable Changed.  Default is 0b.	RW
8:5	<b>Reserved</b>	RO
9	<b>VF Migration Capable</b> – Set by MR-PCIM to indicate to SR-PCIM that VF Migration support is available. VF Migration is possible only if this bit and VF Migration Enabled are both set.  Default is 0b. This field is Read Only Zero if VF Migration Supported is Clear.	RW
15:10	<b>Reserved</b>	RO
31:16	<b>InitialVFs</b> – Number of VFs provided to SR-PCIM and populated by MR-PCIM with MVFs.  This field is meaningful only for PFs. If this Function does not contain an SR-IOV Capability, this field is Read Only Zero.  If VF Mapping is not supported on this PF, this field is Read Only and indicates the number of VFs that were provisioned using Vendor	RW

Bit Location	Register Description	Attributes
	<p>Specific mechanisms.</p> <p>If VF Mapping is supported, the default value of this field is 0000h.</p> <p>The value 0 indicates that no populated VFs are offered to SR-PCIM.</p>	

**Table 4-20: Function Control 3 (10h)**

Bit Location	Register Description	Attributes
15:0	<p><b>Base LVF</b> – Set by MR-PCIM to contain the index of first LVF table entry assigned to this PF.</p> <p>If VF Mapping is not supported, this field is Read Only Zero.</p> <p>The default value of this field is Vendor Specific.</p>	RW
31:16	<p><b>TotalVFs</b> – Total Number of VFs provided to SR-PCIM including populated and non-populated VFs.</p> <p>This field is meaningful only if VF Migration is supported. If VF Migration is not supported, this field is Read Only Zero and the InitialVFs value is returned as TotalVFs to SR-PCIM.</p> <p>The default value of this field is 0000h.</p> <p>MR-PCIM must configure this field to be greater than or equal to InitialVFs.</p> <p>The value 0 indicates that no VFs are offered to SR-PCIM.</p>	RW

#### 4.2.4.3. Function Status (14h)

Setting of any of bits 7:0 of this register Sets the corresponding Function Interrupt Bitmap entry if the Function Interrupt Enable bit is Set. When software clears all of these bits, or clears the Function interrupt Enable bit, the corresponding Function Interrupt Bitmap entry is also cleared.

Table 4-21: Function Status

Bit Location	Register Description	Attributes
0	<b>VC Config Changed</b> – Set when software in a VH changes any of the VC Resource State bits. Default is 0b. If VC Capability Supported is Clear, this field is hardwired 0b.	RW1C
1	<b>MFVC Config Changed</b> – Set when software in a VH changes any of the MFVC Resource State bits. Default is 0b. If MFVC Capability Supported is Clear, this field is hardwired 0b.	RW1C
2	<b>PF Reset Initiated</b> – Set when the PF is Reset. This can be due to a Reset DLLP within the VH or due to software in the VH issuing a Function Level Reset (FLR) to the PF.  This field is Read Only Zero if VF Migration Supported is Clear.	RW1C
3	<b>VF Migration Status</b> – Set when a VF Migration event is triggered for some VF associated with this PF. See Section 3.2.4 for details.  This field is Read Only Zero if VF Migration Supported is Clear.	RW1C
4	<b>VF Enable Changed</b> – Set when SR-PCIM changes VF Enable.	RW1C
5	<b>VF Initialization Pending</b> – Set when VF Migration Capable is Set and either the PF is Reset or VF Enable is Cleared. When Set, access with the VH to VF Configuration Space will return CRS.  This field is Read Only Zero if VF Migration Supported is Clear.  To avoid software within the MR-PCIM software should re-establish a valid initial VF Configuration and Clear this bit within 1 second. See Section 3.2.4.2 for details.	RW1C
7:6	<b>Reserved</b>	RO
8	<b>VF Enabled</b> – MR-IOV visible copy of the SR-IOV VF Enable bit.	RO
9	<b>VF Migration Enabled</b> – MR-IOV visible copy of the SR-IOV VF Migration Enable bit.	RO
15:10	<b>Reserved</b>	RO
31:16	<b>NumVFs</b> – MR-IOV visible copy of the SR-IOV NumVFs value.	RO

#### 4.2.4.4. Function VC to VL Map (18h and 1Ch)

This table maps VCs in a VH to VLs. To preserve PCI Express ordering and flow control independence assumptions, each VC must be assigned to a distinct VL.

This map is only meaningful on the Main BF of the Device. In all other BFs, all fields of this map are Read Only Zero.

**Table 4-22: Function Table VC to VL Map 1 (VC Capability)**

Bit Location	Register Description	Attributes
2:0	<b>VC0 VL Map</b> – Indicates the VL number used for traffic labeled VC0. If VC Capability Supported is Clear or if MaxVL is 0h, this field is Read Only Zero. The default value of this field is 0h.	RW
3	<b>Reserved</b>	RO
4	<b>VC0 VL Map Enable</b> – Indicates that the VC0 VL Map field contains a valid VL number. The default value of this field is 0b. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
7:5	<b>Reserved</b>	RO
10:8	<b>VC1 VL Map</b> – Indicates the VL number used by traffic labeled VC1. This field is Read Only Zero if Num VC Resources Hardware Present is 0. The default value of this field is 1h.	RW
11	<b>Reserved</b>	RO
12	<b>VC1 VL Map Enable</b> – Indicates that the VC1 VL Map field contains a valid VL number. The default value of this field is 0b. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
15:13	<b>Reserved</b>	RO
18:16	<b>VC2 VL Map</b> – Indicates the VL number used by traffic labeled VC2. This field is Read Only Zero if Num VC Resources Hardware Present is 0 or 1. The default value of this field is 2h.	RW
19	<b>Reserved</b>	RO
20	<b>VC2 VL Map Enable</b> – Indicates that the VC2 VL Map field contains a valid VL number. The default value of this field is 0b. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
23:21	<b>Reserved</b>	RO
26:24	<b>VC3 VL Map</b> – Indicates the VL number used by traffic labeled VC3. This field is Read Only Zero if Num VC Resources Hardware Present is 0 to 2. The default value of this field is 3h.	RW
27	<b>Reserved</b>	RO
28	<b>VC3 VL Map Enable</b> – Indicates that the VC3 VL Map field contains a valid VL number. The default value of this field is 0b.	RW

Bit Location	Register Description	Attributes
	Hardware behavior on the 1b to 0b transition of this bit is undefined.	
31:29	<b>Reserved</b>	RO

Table 4-23: Function Table VC to VL Map 2 (VC Capability)

Bit Location	Register Description	Attributes
3:0	<b>VC4 VL Map</b> – Indicates the VL number used by traffic labeled VC4. This field is Read Only Zero if Num VC Resources Hardware Present is 0 to 3. The default value of this field is 4h.	RW
4	<b>Reserved</b>	RO
5	<b>VC4 VL Map Enable</b> – Indicates that the VC4 VL Map field contains a valid VL number. The default value of this field is 0b. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
7:6	<b>Reserved</b>	RO
10:8	<b>VC5 VL Map</b> – Indicates the VL number used by traffic labeled VC5. This field is Read Only Zero if Num VC Resources Hardware Present is 0 to 4. The default value of this field is 5h.	RW
11	<b>Reserved</b>	RO
12	<b>VC5 VL Map Enable</b> – Indicates that the VC5 VL Map field contains a valid VL number. The default value of this field is 0b. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
15:13	<b>Reserved</b>	RO
18:16	<b>VC6 VL Map</b> – Indicates the VL number used by traffic labeled VC6. This field is Read Only Zero if Num VC Resources Hardware Present is 0 to 5. The default value of this field is 6h.	RW
19	<b>Reserved</b>	RO
20	<b>VC6 VL Map Enable</b> – Indicates that the VC6 VL Map field contains a valid VL number. The default value of this field is 0b. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
23:21	<b>Reserved</b>	RO
26:24	<b>VC7 VL Map</b> – Indicates the VL number used by traffic labeled VC7. This field is Read Only Zero if Num VC Resources Hardware Present is 0 to 6. The default value of this field is 7h.	RW
27	<b>Reserved</b>	RO
28	<b>VC7 VL Map Enable</b> – Indicates that the VC7 VL Map field contains a valid VL number. The default value of this field is 0b. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
31:29	<b>Reserved</b>	RO

#### 4.2.4.5. Function VC Resource (20h to 2Ch)

These fields return data from the VC Capability of the associated Type 1 Configuration Header. They allow MR-PCIM software to track the enabling and mapping of VCs with each VH.

VC Resource fields for resource numbers above Num VC Resource Hardware Present are not implemented and return 0 when read. VC Resource fields for resource numbers above Extended VC Count are Undefined.

VC Resource State 0 is located at offset 1Ch; VC Resource 1 is located at offset 1Eh; etc. Fields within VC Resource State are described in [Table 4-25](#).

Deleted: Table 4-25

**Table 4-24: Function Table VC Resource State**

Bit Location	Register Description	Attributes
0	<b>VC Enabled</b> – This field tracks the VC Enabled bit set by software operating in the VH. Per the PCI Express specification, VC Resource 0 is always enabled and thus VC Enabled for VC Resource 0 is always set.	RO
1	<b>VC Negotiation Pending</b> – This field tracks the VC Negotiation Pending bit view in the VH. This bit is Set when VC Enabled is Set and either no VL has been mapped to this VC (associated VL Map Enable bit is Clear) or Flow Control negotiation has not completed on the mapped VH and VL.	RO
3:2	<b>Reserved</b>	RO
6:4	<b>VC ID</b> – This field tracks the VC ID field set by software operating in the VH. Per the PCI Express specification, VC ID for VC Resource 0 is always 0.	RO
7	<b>Reserved</b>	RO
15:8	<b>TC to VC Map</b> – This field tracks the TC to VC Map field set by software operating in the VH. Per the PCI Express specification, bit 0 of this field is fixed and the remaining bits may be set by software.	RO

#### 4.2.4.6. Function Table MFVC Resource Status (30h to 3Ch)

These fields return data from the MFVC Capability of the associated Type 1 Configuration Header. They allow MR-PCIM software to track the enabling and mapping of VCs with each VH.

VC Resource fields for resource numbers above Num MFVC Resource Hardware Present are not implemented and return 0 when read. VC Resource fields for resource numbers above Extended MFVC Count are Undefined.

VC Resource State 0 is located at offset 30h; VC Resource 1 is located at offset 32h; etc. Fields within VC Resource State are described in [Table 4-25](#).

Deleted: Table 4-25

**Table 4-25: VH Table MFVC Resource State**

Bit Location	Register Description	Attributes
0	<b>VC Enabled</b> – This field tracks the VC Enabled bit set by software operating in the VH. Per the PCI Express specification, VC Resource 0 is always enabled and thus VC Enabled for VC Resource 0 is always set.	RO
1	<b>VC Negotiation Pending</b> – This field tracks the VC Negotiation Pending bit view in the VH. This bit is Set when VC Enabled is Set and either no VL has been mapped to this VC (associated VL Map Enable bit is Clear) or Flow Control negotiation has not completed on the mapped VH and VL.	RO
3:2	<b>Reserved</b>	RO
6:4	<b>VC ID</b> – This field tracks the VC ID field set by software operating in the VH. Per the PCI Express specification, VC ID for VC Resource 0 is always 0.	RO
7	<b>Reserved</b>	RO
15:8	<b>TC to VC Map</b> – This field tracks the TC to VC Map field set by software operating in the VH. Per the PCI Express specification, bit 0 of this field is fixed and the remaining bits may be set by software.	RO

#### 4.2.4.7. Function Interrupt Bitmap (minus 20h)

The Function Interrupt Status bitmap precedes the Function Table. It is always 32 bytes (supporting switches with a maximum of 256 ports).

Bits in this table are Read Only. A bit is Set to indicate the Function has an interrupt pending and Clear otherwise. These Interrupt Status bits are cleared either by clearing the appropriate Function Interrupt Pending bit or by masking the interrupt using the Function Interrupt Enable.

An MSI Interrupt is requested on any zero to one transition of any of these bits.

### 4.2.5. Misc. Device Configuration Space Requirements

#### 4.2.5.1. BIST (Device)

BIST remains optional in MR-IOV. The results of invoking BIST in any non-BF Function must not affect any other VH.

The results are undefined if software invokes BIST in a BF when any VC to VL Map Enable bit in any Function Table entry of any BF of the Device is Set..

## 4.3. Switch Configuration Space

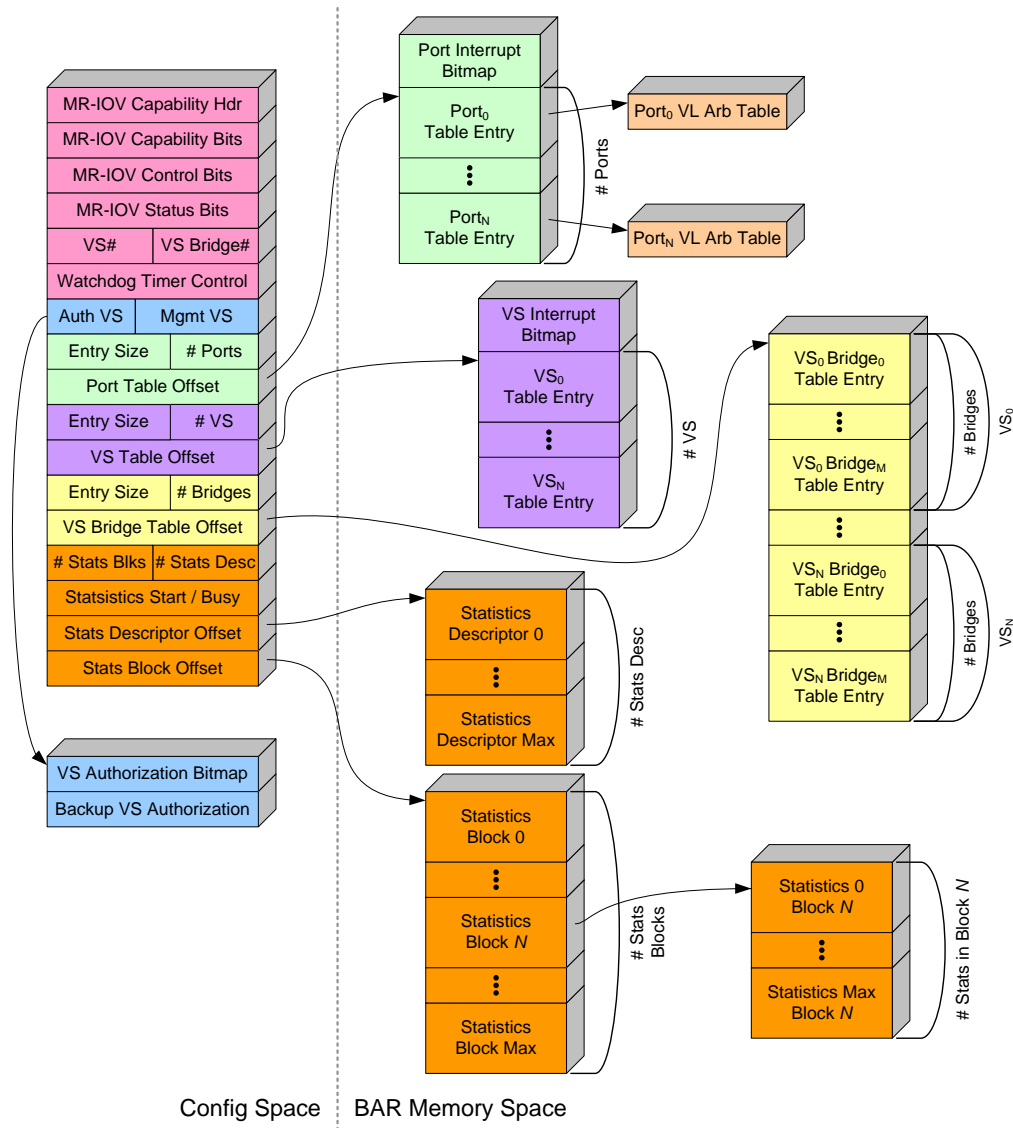
For managing MR-IOV Switches, MR-PCIM must use the Switch Type 1 Headers to manage the Switch. Configuration controls are associated with:

- ☐ Entire component
- ☐ Physical Port of the component
- ☐ Virtual Switch within the component
- ☐ PCI-to-PCI Bridges within each Virtual Switch

There are nine tables provided by the Switch. These tables are located using the MR-IOV Capability block located in the Type 1 Configuration Space of upstream P2P Bridge(s). An overview of the tables is show in [Figure 4-5](#).

Deleted: Figure 4-5





**Figure 4-5: Switch Mapping Tables**

- ❑ The MR-IOV Capability contains the controls that apply to the entire Switch. It also contains BAR relative offsets to tables located in BAR Memory Space and associated table size information.
- ❑ The VS Authorization Bitmap contains one bit for each Virtual Switch. If the corresponding bit is Set, the associated VS is authorized and can be used to manage the MR switch.

- ❑ The Port Table contains an entry for every Port on the Switch. The table may be sparse (i.e. contain unused table entries) for hardware implementation flexibility. The Port Table contains the fields that control the physical link. The Port table also points to the optional VL Arbitration Table. Preceding the Port Table is a 256 bit Port Interrupt Summary.
- ❑ The optional VL Arbitration Table supports controlling the arbitration between Virtual Links for access to a Port. This table is modeled after the VC Arbitration Table in PCI Express.
- ❑ The VS Table contains an entry for every Virtual Switch in the MR Switch. The table may be sparse for hardware implementation flexibility. Preceding the VS Table is a 256 bit VS Interrupt Summary.
- ❑ The VS Bridge Table contains an entry for every PCI-to-PCI bridge in every Virtual Switch. This table is a two dimensional array indexed by VS number and by P2P Bridge number. Within a VS, the first entry corresponds to the upstream P2P Bridge and the remaining entries are the downstream bridge(s). This table may also be sparse but the upstream entry of each VS must be present if the associated VS is present. No entries are present in this table unless the associated VS table is also present.
- ❑ The optional Statistics Descriptor Table contains descriptions of the varieties of performance counters and statistics information supported. This table is read only and contains one entry for each counter style supported by the component.
- ❑ The optional Statistics Block Table contains an entry for every block of related statistics counters. Each entry contains controls for the block and an offset to the array of counters.
- ❑ The optional Statistics Counter Table contains the actual counters and sampled values. There is one table for each Statistics Block. The offset and size of this table is contained in the associated Statistics Block Table.

These tables must be visible in the Upstream P2P Bridge of all authorized VHs as. A subset of the MR-IOV Capability (but none of the other tables) may optionally be present in the Upstream P2P Bridges of non-authorized VHs. This subset MR-IOV Capability must be present in VH 0 Upstream P2P Bridges that are attached to MR aware Root Ports.

#### 4.3.1. Switch MR-IOV Extended Capability

[Figure 4-6](#) shows the Switch MR-IOV Extended Capability structure in more detail.

Deleted: Figure 4-6

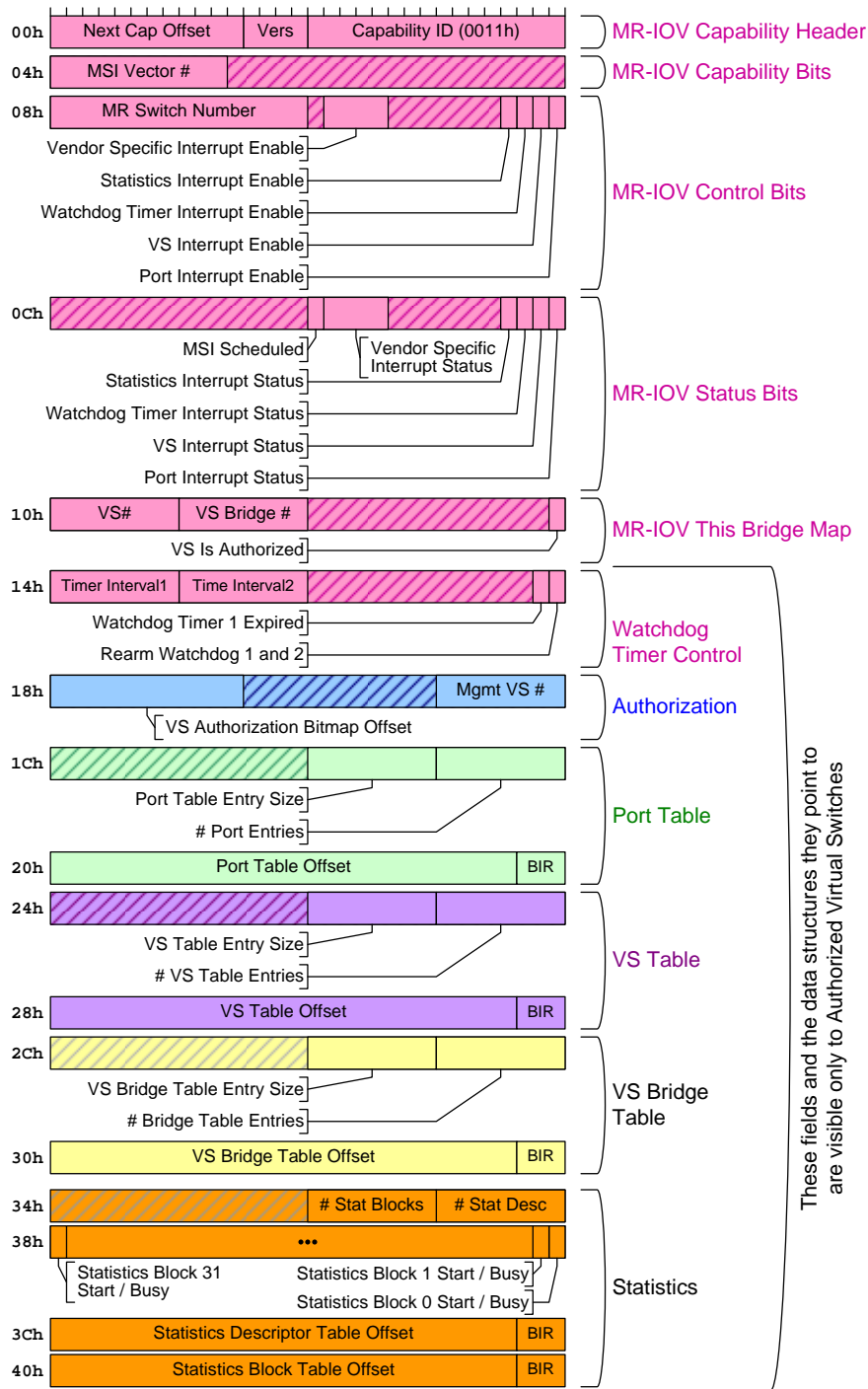


Figure 4-6: Switch MR-IOV Capability Diagram

4.3.1.1. Switch MR-IOV Extended Capability Header (00h)

Table 4-26 defines the Switch MR-IOV Extended Capability Header. The Capability ID for the Switch MR-IOV Extended Capability is 0011h.

Deleted: Table 4-26

Table 4-26: Switch MR-IOV Extended Capability Header

Bit Location	Register Description	Attributes
15:0	<b>PCI Express Extended Capability ID</b> – This field is a PCI-SIG defined ID number that indicates the nature and format of the Extended Capability.  The Extended Capability ID for the MR-IOV Extended Capability is 0011h.	RO
19:16	<b>Capability Version</b> – This field is a PCI-SIG defined version number that indicates the version of the Capability structure present.  Must be 1h for this version of the specification.	RO
31:20	<b>Next Capability Offset</b> – This field contains the offset to the next PCI Express Capability structure or 000h if no other items exist in the linked list of Capabilities.  This offset is relative to the beginning of PCI compatible Configuration Space and thus must always be either 000h (for terminating list of Capabilities) or greater than 0FFh.	RO

4.3.1.2. Switch MR-IOV Capability (04h)

Table 4-27: Switch MR-IOV Capability Bits

Bit Location	Register Description	Attributes
21:0	<b>Reserved</b>	RO
31:21	<b>MSI Vector Number</b> – This field indicates the MSI Vector number used to signal MR-IOV events within this Switch. This value may change based on whether MSI or MSI-X is enabled.	RO

4.3.1.3. *Switch MR-IOV Control (08h)*

Table 4-28: Switch MR-IOV Control Bits

Bit Location	Register Description	Attributes
0	<b>Port Interrupt Enable</b> – Enables delivery Port Interrupts. See the Port Table for details. This bit is implemented per MR-IOV capability. This bit is Read Only Zero unless this VS is Authorized. The default value of this field is 0b.	RW
1	<b>VS Interrupt Enable</b> – Enables delivery of VS Interrupts. See the VS Table for details. This bit is implemented per MR-IOV capability. This bit is Read Only Zero unless this VS is Authorized. The default value of this field is 0b.	RW
2	<b>Watchdog Timer Interrupt Enable</b> – Enables delivery of Watchdog Timer Interrupts. This bit is implemented per MR-IOV capability. This bit is Read Only Zero unless this VS the is Authorized or is Authorized in the Shadow Authorization bitmap. The default value of this field is 0b.	RW
10:3	<b>ReservedP</b>	RO
14:11	<b>Vendor Specific Interrupt Enable</b> – These bits enable delivery of a variety of Vendor Specific Interrupts.  All Vendor Specific Interrupts are optional. Unimplemented Interrupt Enable bits are Read Only Zero. Different Ports may implement different Vendor Specific Interrupts.  Implementation and usage of these bits may vary based on whether the VS is Authorized. The default value of this field is 0000b.	RW
15	<b>ReservedP</b>	RO
31:16	<b>MR Switch Number</b> – Scratchpad register used by MR-PCIM in detecting loops during Topology Enumeration. This field is Read Only unless this VS is Authorized.	RW

#### 4.3.1.4. Switch MR-IOV Status (0Ch)

**Table 4-29: Switch MR-IOV Status Bits**

Bit Location	Register Description	Attributes
0	<b>Port Interrupt Status</b> – Indicates that a Port Interrupt is pending. This bit is Read Only and indicates that some bit in the Port Interrupt Bitmask is Set. This bit is Read Only Zero if this VS is not Authorized.	RO
1	<b>VS Interrupt Status</b> – Indicates that a VS Interrupt is pending. This bit is Read Only and indicates that some bit in the VS Interrupt Bitmask is Set. This bit is Read Only Zero if this VS is not Authorized.	RO
2	<b>Watchdog Timer Interrupt Status</b> – Indicates that the watchdog timer expired. This bit is implemented per MR-IOV capability.	RW1C
10:3	<b>ReservedZ</b>	RO
14:11	<b>Vendor Specific Interrupt Status</b> – These bits indicate that one of a variety of Vendor Specific Interrupt is pending.  All Vendor Specific Interrupts are optional. Unimplemented Interrupt Status bits are Read Only Zero.  Bits in this field are either RW1C or RO. If RO, clearing the interrupt status involves Vendor Specific mechanisms.  Implementation and usage of bits may vary based on whether the VS is Authorized.	RW1C / RO
15	<b>MSI Scheduled</b> – Set when an MSI has been requested. If Set, subsequent MSIs are suppressed. If Clear, any enabled interrupt will cause an MSI to be scheduled (and this bit to be Set). Default is 0b.	RW1C
31:16	<b>ReservedZ</b>	RO

#### 4.3.1.5. MR-IOV This Bridge Map (10h)

This Read Only register returns the VS number and P2P Bridge number within the VS of this type 1 config header. In the upstream P2P Bridge of a VS, it also indicates the Authorization status of the VS.

**Table 4-30: Switch MR-IOV This Bridge Map**

Bit Location	Register Description	Attributes
0	<b>VS is Authorized</b> – Indicates that the VS is authorized to manage the switch. See Authorization Control for details.	RO
15:1	<b>ReservedZ</b>	RO
23:16	<b>VS Bridge Number</b> – Indicates the P2P Bridge number within the indicated VS of this Type 1 config space. This value can be used to locate the associated VS Bridge Table Entry	RO
31:24	<b>VS Number</b> – Indicates the VS number of this Type 1 config space. This value can be used to locate the associated VS Table and VS Bridge Table Entries.	RO

#### 4.3.1.6. Watchdog Timer Control (14h)

The Watchdog Timer is used ensure that a backup MR-PCIM can recover and take over from the primary MR-PCIM. One key area is the situation where discovery has failed and the Initial MR-PCIM has altered the switch configuration(s) such that the backup MR-PCIM can't manage a switch (e.g. the Backup MR-PCIM VS could have been de-authorized, some inter-switch link directions could be configured inappropriately, etc.).

Two Watchdog Timers are provided:

- ❑ Timer 1 supports sending an interrupt, reauthorizing selected Virtual Switches and resetting the Link Direction on selected Ports.
- ❑ Timer 2 supports a complete reset of the switch. This can be used as a “fall back” mechanism in the case that Timer 1 was not able to reconfigure.

Table 4-31: Switch MR-IOV Authorization Control

Bit Location	Register Description	Attributes
0	<b>Rearm Watchdog 1 and 2</b> – Writing 1 causes both Watchdog Timers to start counting. Writing 0 has no effect. Reads as zero.	RW
1	<b>Timer 1 Expired</b> – Returns 1 if Watchdog Timer 2 has expired. Returns 0 if the timer is disabled or is still counting.	RO
15:2	<b>ReservedZ</b>	RO
23:16	<b>Watchdog Timer Interval 2</b> – Determines the duration of the Watchdog Timer 2. The timer expires if this period of time elapses before software restarts the watchdog timer.  If the timer expires, the MR Switch is returned to its Initial Power-On Condition (e.g. parameters are reloaded based on straps, EEPROM, etc.).  Encoding for this field is TBD. One encoding is used to indicate the timer is disabled and will never expire.	RW
31:24	<b>Watchdog Timer Interval 1</b> – Determines the duration of the Watchdog Timer 1. The timer expires if this period of time elapses before software restarts the watchdog timer.  If the timer expires, an interrupt is signaled, the backup Authorization Bitmap is copied to the Authorization Bitmap, and for all Ports, the Backup Link Direction Control is copied to the Link Direction Control.  Encoding for this field is TBD. One encoding is used to indicate the timer is disabled and will never expire.	RW

#### 4.3.1.7. Authorization (18h)

This register determines which Virtual Switches are authorized to manage the MR Switch. It also indicates which VS receives “route to MR-PCIM” messages initiated by this switch.

Table 4-32: Switch MR-IOV Authorization Control

Bit Location	Register Description	Attributes
7:0	<b>Management VS</b> – Indicates which VS is considered the primary management VS. The active MR-PCIM is running above the Root Port at the top of the hierarchy containing this VS.  The indicated VS is automatically authorized independent of the state of the corresponding VS Authorization Bitmap entry.  This field is Read Only Zero if this VS is not Authorized.	RW
19:8	<b>ReservedP</b>	RO
31:20	<b>VS Authorized Bitmap Offset</b> – This field contains the offset to the Authorization Bitmap. This offset is relative to the beginning of PCI compatible Configuration Space. See section 4.3.2 for details.	RO



#### 4.3.1.8. Port Table Entry Size / Num Port Entries (1Ch)

**Table 4-33: Switch Port Table Sizes**

Bit Location	Register Description	Attributes
7:0	<b>Num_Port_Table_Entries</b> – Returns the number of entries in the Port Table.	RO
15:8	<b>Port_Table_Entry_Size</b> – Returns the size of a Port Table Entry in DWORDs. For the current version of this specification this value must be 15h or larger. Implementations may use larger values to simplify address arithmetic.	RO
31:16	<b>ReservedZ</b>	RO

The total size of the Port Table (in bytes) is:

$$\text{Num\_Port\_Table\_Entries} * \text{Port\_Table\_Entry\_Size} * 4$$

#### 4.3.1.9. Port Table Offset (20h)

**Table 4-34: Switch Port Table Offset**

Bit Location	Register Description	Attributes									
2:0	<p><b>Port Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's Port Table into Memory Space.</p> <p><b>BIR Value Base Address register</b></p> <table> <tr> <td>0</td><td>BAR0</td><td>10h</td></tr> <tr> <td>1</td><td>BAR1</td><td>14h</td></tr> <tr> <td>2..7</td><td>Reserved</td><td></td></tr> </table> <p>For a 64-bit Base Address register, the BIR indicates the lower DWORD.</p>	0	BAR0	10h	1	BAR1	14h	2..7	Reserved		RO
0	BAR0	10h									
1	BAR1	14h									
2..7	Reserved										
31:3	<b>Port Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the Port Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.	RO									

The Port Table starts at the Port Table Offset. The Port Interrupt Bitmap immediately precedes the Port Table (i.e. it starts 32 bytes before the Port\_Table\_Offset).

#### 4.3.1.10. VS Table Entry Size / Num VS Table Entries (24h)

**Table 4-35: Switch VS Table Sizes**

Bit Location	Register Description	Attributes
7:0	<b>Num_VS_Table_Entries</b> – Returns the number of entries in the VS Table.	RO
15:8	<b>VS_Table_Entry_Size</b> – Returns the size of a VS Table Entry in DWORDs. For the current version of this specification this value must be 3h or larger (more if some VS supports more than 32 Bridges). Implementations may use larger values to simplify address arithmetic.	RO
31:16	<b>Reserved</b>	RO

The total size of the VS Table (in bytes) is:

$$\text{Num\_VS\_Table\_Entries} * \text{VS\_Table\_Entry\_Size} * 4$$

#### 4.3.1.11. VS Table Offset (28h)

**Table 4-36: Switch VS Table Offset**

Bit Location	Register Description	Attributes									
2:0	<p><b>VS Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's VS Table into Memory Space.</p> <p><b>BIR Value Base Address register</b></p> <table> <tr> <td>0</td><td>BAR0</td><td>10h</td></tr> <tr> <td>1</td><td>BAR1</td><td>14h</td></tr> <tr> <td>2..7</td><td>Reserved</td><td></td></tr> </table> <p>For a 64-bit Base Address register, the BIR indicates the lower DWORD.</p>	0	BAR0	10h	1	BAR1	14h	2..7	Reserved		RO
0	BAR0	10h									
1	BAR1	14h									
2..7	Reserved										
31:3	<b>VS Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the VS Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.	RO									

The VS Table starts at the VS Table Offset. The VS Interrupt Bitmap immediately precedes the VS Table (i.e. it starts 32 bytes before the VS\_Table\_Offset).

#### 4.3.1.12. VS Bridge Table Entry Size / Num VS Bridge Table Entries per VS (2Ch)

Table 4-37: Switch VS Bridge Table Sizes

Bit Location	Register Description	Attributes
7:0	<b>Num_Bridge_Table_Entries</b> – Returns the number of Bridge entries in the VS Bridge Table associated with each VS.	RO
15:8	<b>Bridge_Table_Entry_Size</b> – Returns the size of a VS Bridge Table Entry in DWORDs. For the current version of this specification, this value must be at least 8. Implementations may use larger values to simplify address arithmetic.	RO
31:16	<b>Reserved</b>	RO

The total size (in bytes) of the VS Bridge Table is:

$$\text{Num\_Bridge\_Table\_Entries} * \text{Num\_VS\_Table\_Entries} * \text{Bridge\_Table\_Entry\_Size} * 4$$

Note: Num\_Bridge\_Table\_Entries reflects the size of the table and thus is the maximum size across all Virtual Switches. Not all VS Bridge Table Entries need be present (see Section 4.3.6.1).

#### 4.3.1.13. VS Bridge Table Offset (30h)

Table 4-38: Switch VS Bridge Table Offset

Bit Location	Register Description	Attributes									
2:0	<p><b>VS Bridge Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's VS Bridge Table into Memory Space.</p> <p><b>BIR Value Base Address register</b></p> <table> <tr> <td>0</td><td>BAR0</td><td>10h</td></tr> <tr> <td>1</td><td>BAR1</td><td>14h</td></tr> <tr> <td>2..7</td><td>Reserved</td><td></td></tr> </table> <p>For a 64-bit Base Address register, the BIR indicates the lower DWORD.</p>	0	BAR0	10h	1	BAR1	14h	2..7	Reserved		RO
0	BAR0	10h									
1	BAR1	14h									
2..7	Reserved										
31:3	<b>VS Bridge Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the VS Bridge Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.	RO									

The VS Bridge Table Entry associated with Bridge 0 of VS *N* immediately follows the last Bridge Table Entry associated with VS *N-1*.

#### 4.3.1.14. *Statistics Capability and Control (30h to 3Ch)*

Device and Switch Statistics related fields are described in section **TBD**.

#### 4.3.2. Switch VS Authorization Bitmap

The VS Authorization Bitmap and Backup VS Authorization Bitmap are located in Configuration space at the byte offset indicated by the VS Authorization Bitmap Offset. These tables must be located in extended Configuration Space (i.e. the offset must be greater than FFh).

Both tables contain Num\_VS\_Table\_Entries bits. Bit 0 of the first byte corresponds to VS0, bit 1 corresponds to VS1, etc.

A VS is authorized if either the corresponding bit in the VS Authorization Bitmap is Set or if the VS is the Management VS.

Only transactions from Authorized VSs are allowed to access memory space MR-IOV tables of this Switch.

Only transactions from Authorized VSs are allowed to access certain fields Config space of the MR-IOV Extended Capability.

Both Authorization Bitmaps are Read Only Zero if this VS is not Authorized.

Note: The bitmap entry for the Management VS is read/write but its value does not affect Authorization state of that VS. If the “old” Management VS is to remain Authorized after a transition to a “new” Management VS, software should Set the “old” VS this before changing Management VS to ensure that Authorization remains seamless.

The Backup VS Authorization Bitmap is copied to the VS Authorization Bitmap when Watchdog Timer 1 expires. The Backup VS Authorization Bitmap values have no other effect.

#### 4.3.3. Switch Port Table

The Port Table contains up to 256 Port Table Entries. This table is located in memory starting at a location determined by the Port Table Offset field in the MR-IOV Capability.

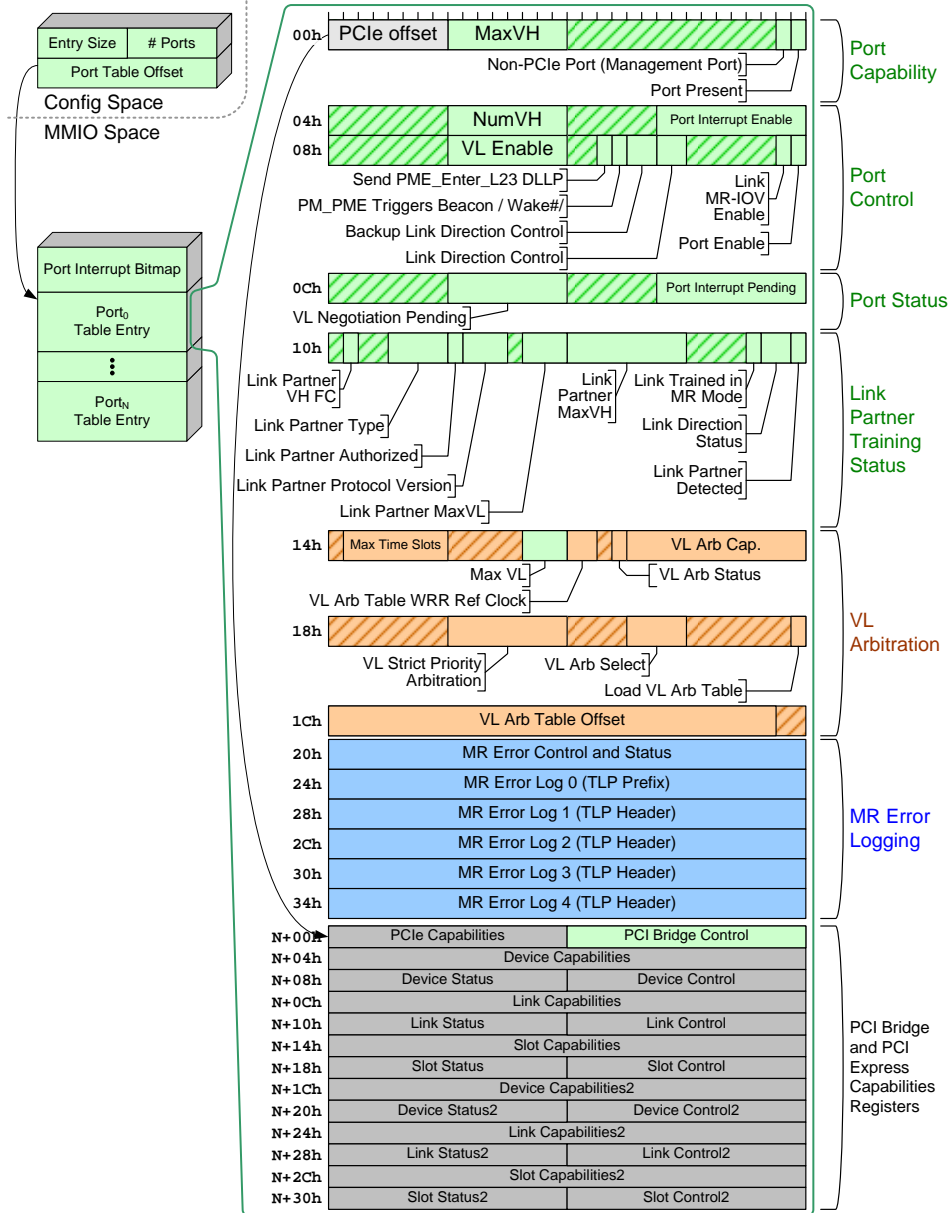


Figure 4-7: Switch Port Table

4.3.3.1. Port Capability (00h)

Table 4-39: Switch Port Capability

Bit Location	Register Description	Attributes
0	<b>Port Present</b> – Indicates that this Port is present on the Switch.	RO
1	<b>Non-PCIe Switch Management Port</b> – Indicates that this is Vendor Specific non-PCI Express Port used to manage a switch (see section 3.1.1.3).  Certain bits in the Port Entry are no longer meaningful. <b>TODO: Specify these bits.</b>	RO
15:2	<b>Reserved</b>	RO
23:16	<b>MaxVH</b> – Maximum number of VHs supported on this Device	RO
31:24	<b>PCIe Offset</b> – DWORD Offset to the PCIe Capabilities section of the Port Table. For the current version of the specification, this must be at least 0Eh (i.e. 38h divided by 4).	RO

#### 4.3.3.2. Port Control (04h and 08h)

Table 4-40: Switch Port Control1

Bit Location	Register Description	Attributes
0	<b>Port DL_Up Interrupt Enable</b> – When both this bit and the Port DL_Up Interrupt Pending bit are set, the Port interrupt Status bit associated with this Port is Set and a Port Interrupt is requested.	RW
1	<b>Port DL_Down Interrupt Enable</b> – When both this bit and the Port DL_Down Interrupt Pending bit are set, the Port Interrupt Status bit associated with this Port is Set and a Port Interrupt is requested.	RW
2	<b>Port PME Turn Off Interrupt Enable</b> – When both this bit and the Port PME Turn Off Interrupt Pending bit are set, the Port Interrupt Status bit associated with this Port is Set and a Port Interrupt is requested.	RW
3	<b>Link Retrain Interrupt Enable</b> – When both this bit and the Link Retrain Interrupt Pending bit are set, the Port Interrupt Status bit associated with this Port is Set and a Port Interrupt is requested.	RW
4	<b>Beacon / WAKE# Interrupt Enable</b> – When both this bit and the Beacon / WAKE# Interrupt Pending bit are set, the Port Interrupt Status bit associated with this Port is Set and a Port Interrupt is requested.  Beacon / WAKE# support is optional. If not supported, this bit is Read Only Zero.	RW
5	<b>MR Error Uncorrectable Interrupt Enable</b> – When both this bit and the MR Uncorrectable Error Interrupt Pending bit are set, the Port Interrupt Status bit associated with this port is Set and a Port Interrupt is requested.	RW
6	<b>MR Advisory Error Interrupt Enable</b> – When both this bit and the MR Advisory Error Interrupt Pending bit are set, the Port Interrupt Status bit associated with this Port is Set and a Port Interrupt is requested.	RW
7	<b>Physical Hot-Plug Interrupt Enable</b> – When both this bit and the Physical Hot-Plug Interrupt Pending bit are set, the port Interrupt Status bit associated with this Port is Set and a Port Interrupt is requested.	RW
15:8	<b>Reserved</b>	RO
23:16	<b>NumVH</b> – Indicates the number of VHs enabled. This value must be less than or equal to MaxVH. The default value of this field is Vendor Specific.  This value may be used by the Switch to optimize resource usage.	RW
31:24	<b>Reserved</b>	RO

**Table 4-41: Switch Port Control2**

Bit Location	Register Description	Attributes
0	<b>Port Enable</b> – Set by MR-PCIM to indicate that it wishes to use the Port.	RW
1	<b>Link MR-IOV Enable</b> – If Set, the Link will attempt to negotiate to use the MR-IOV enhanced protocol. If Clear, the link will not attempt to use MR-IOV and will thus train the link in Base PCIe mode. Read Only if the Link is up as indicated by Link Direction Status.	RW
7:2	<b>Reserved</b>	RO
9:8	<b>Link Direction Control</b> – Controls how the Link should train. Values are: 0   Upstream Switch Port 1   Downstream Switch Port 2   Cross-Link (if supported) 3   Don't Train, keep Link down	RW
11:10	<b>Backup Link Direction Control</b> – The Backup Link Direction Control value is copied to the Link Direction Control field when Watchdog Timer 1 expires. The Backup Link Direction Control value has no other effect.	RW
12	<b>PM_PME Triggers Beacon / WAKE#</b> – If Set, the automatic triggering of Beacon / WAKE# on reception of a Beacon, WAKE# or PM_PME message at this port is suppressed. See Section 7.6 for details.	RW
13	<b>Send PME_Enter_L23 DLLP</b> – If software writes 1b to this bit and this Port's Link Direction Status indicates Upstream Switch Port, initiate the PME_Enter_L23 handshake to power down the Link.  Writing 0 to this bit has no effect. If Link Direction Status is not Upstream Switch Port, writing any value to this bit has no effect.	Write 1 to Send
15:14	<b>Reserved</b>	RO
23:16	<b>VL Enable</b> –This bit, when Set, enables a Virtual Link – This bit, when Set, enables a Virtual Link (see note 1 for exceptions). The Virtual Link is disabled when this bit is cleared.  Software must use the VL Negotiation Pending bit to check whether the VL negotiation is complete.  Default value of this bit is 1b for the first VL and is 0b for other VLs.  Notes: 1. This bit is hardwired to 1b for the VL0, i.e., writing to this bit has no effect for VL0. 2. To enable a Virtual Link, the VL Enable bits for that Virtual Link must be Set in both components on a Link. 3. To disable a Virtual Link, the VL Enable bits for that Virtual Link must be cleared in both components on a Link. 4. Software must ensure that no traffic is using a Virtual Link at the time it is disabled.	RW



Bit Location	Register Description	Attributes
	5. Software must fully disable a Virtual Link in both components on a Link before re-enabling the Virtual Link.	
31:24	Reserved	RO

#### 4.3.3.3. Port Status (0Ch)

Table 4-42: Switch Port Status

Bit Location	Register Description	Attributes
0	<b>Port DL_Up Interrupt Pending</b> – Set the Link enters DL_Up. This bit is Cleared when software writes 1b.	RW1C
1	<b>Port DL_Down Interrupt Pending</b> – Set the Link enters DL_Down. This bit is Cleared when software writes 1b.	RW1C
2	<b>Port PME Turn Off Interrupt Pending</b> – Set when all VS Bridges mapped to this Port and associated with a non-Authorized VS have their PME Turn Off State Set indicating the completion of the PME Turn Off handshake.  This bit is Cleared when software writes 1b.	RW1C
3	<b>Link Retrain Interrupt Pending</b> – Set when the link retrains. Can be used to monitor link health.	RW1C
4	<b>Beacon / WAKE# Interrupt Pending</b> – Set when the link detects a Beacon or WAKE# event. Beacon / WAKE# support is optional. If not supported, this bit is Read Only Zero. When supported, it is form factor specific whether Beacon or WAKE# is used.	RW1C
5	<b>MR Uncorrectable Error Interrupt Pending</b> – Set when the link detects an MR Uncorrectable Error and Sets one of the MR Error Status bits.	RW1C
6	<b>MR Advisory Error Interrupt Pending</b> – Set when the link detects an MR Advisory Error and Sets one of the MR Error Status bits.	RW1C
7	<b>Physical Hot-Plug Interrupt Pending</b> – Set when the Physical Hot-Plug controller indicates that software should be notified.	RW1C
15:8	<b>Reserved</b>	RO
23:16	<b>VL Negotiation Pending</b> – These bits indicate whether Virtual Link Negotiation for some VL is in pending state.  The value of this bit is defined only when the Link is in the DL_Active state and the Virtual Link is enabled (its VL Enable bit is Set).  When this bit is Set by hardware, it indicates that the VL resource has not completed the process of negotiation. This bit is cleared by hardware after the VL negotiation is complete (on exit from the MR FC_INIT2 state on the VL).  Before using a Virtual Link, software must check whether the VL Negotiation Pending bits for that Virtual Link are Clear in both components on the Link.	RO
31:24	<b>Reserved</b>	RO

#### 4.3.3.4. Link Partner Training Status (10h)

Table 4-43: Switch Link Partner Training Status

Bit Location	Register Description	Attributes
0	<b>Link Partner Detected</b> – Set to indicate that a PCI Express component was detected at the remote end of the one or more lanes of the link.	RO
2:1	<b>Link Direction Status</b> – Indicates whether (and how) the link trained. Values are: 0 Upstream Switch Port 1 Downstream Switch Port 2 Reserved 3 Link Down	RO
3	<b>Link Partner is MR</b> – Link was successfully brought up in MR-IOV mode.	RO
7:4	<b>Reserved</b>	RO
31:8	<b>MR Init DLLP Bits</b> – These bits were captured during link training from bytes 1 to 3 of the MR Init DLLP that was sent by the Link Partner. These bits are not meaningful unless Link Partner is MR is Set. These fields allow MR-PCIM to know some information about the Link Partner without needing to read Configuration space (which is not possible if the Link Partner is an MR Root Port since Config requests cannot flow Upstream). The bits are further described <a href="#">Table 4-44</a> below.	RO

Deleted: Table 4-44

**Table 4-44: Switch Link Partner Training Status – MRInit DLLP Bits**

<b>Bit Location</b>	<b>Register Description</b>	<b>Attributes</b>
15:8	<b>Link Partner MaxVH</b> – Maximum number of VHs that the Link Partner can support.	RO
18:16	<b>Link Partner MaxVL</b> – Maximum number of VLs that the Link Partner can support.	RO
19	<b>Reserved</b>	RO
22:20	<b>Link Partner Protocol Version</b> – MR-IOV Protocol version supported by the Link Partner. For this version of the specification, this is the value 1h.	RO
23	<b>Link Partner was Authorized</b> – If the Link Partner is an MR Switch, this bit indicates that at the time of Link Training the VS associated with VH0 of this link was allowed to manage the switch (authorization can be revoked so this may no longer be accurate). This corresponds to the VS is Authorized bit in the Link Partner's MR-IOV Capability.	RO
27:24	<b>Link Partner Type</b> – Indicates what kind of PCI Express Device is present as Function 0 of the Link Partner. Encoding is identical to the Device/Port Type field in the PCI Express Capabilities (Offset 02h, Bits 7:4).	RO
29:28	<b>Reserved</b>	RO
30	<b>Link Partner VH FC</b> – If Set, indicates the Link Partner supports per-VH and Per-VL Flow Control. If Clear, indicates that the Link Partner supports only per-VL Flow Control.	RO
31	<b>Reserved</b>	RO

#### 4.3.3.5. VL Arbitration Capability and Status (14h)

Table 4-45: Switch VL Arbitration Capability and Status

Bit Location	Register Description	Attributes
11:0	<p><b>VL Arbitration Capability</b> – Indicates the types of VL Arbitration supported by the Port. This field is valid for all Functions that report a Low Priority Extended VC Count field greater than 0. For all other Functions, this field must be hardwired to 00h.</p> <p>Each bit location within this field corresponds to a VC Arbitration Capability defined below. When more than 1 bit in this field is Set, it indicates that the Port can be configured to provide different VC arbitration services. Defined bit positions are:</p> <p>Bit 0      Hardware fixed arbitration scheme, e.g., Round Robin</p> <p>Bit 1      Weighted Round Robin (WRR) arbitration with 32 phases</p> <p>Bit 2      WRR arbitration with 64 phases</p> <p>Bit 3      WRR arbitration with 128 phases</p> <p>Bit 4      Time-based WRR with 128 phases</p> <p>Bit 5      WRR Arbitration with 256 phases</p> <p>Bits 6-10   Reserved</p> <p>Bit 10     Vendor Defined VL Arbitration Scheme</p>	RO
12	<p><b>VL Arb Status</b> – This bit indicates the coherency status of the VL Arbitration Table. This bit is valid only when the VL Arbitration Table is used.</p> <p>This bit is Set by hardware when any entry of the VL Arbitration Table is written to by software. This bit is cleared by hardware when hardware finishes loading values stored in the VL Arbitration Table after software sets the Load VL Arbitration Table bit.</p> <p>Default value of this bit is 0b.</p>	RO
13	<b>Reserved</b>	RO
15:14	<p><b>Reference Clock</b> – Indicates the reference clock for Virtual Links that support time-based WRR VL Arbitration. This field is valid only if time-based WRR is supported.</p> <p>Defined encodings are:</p> <p>00b              100 ns reference clock</p> <p>01b – 11b      Reserved</p>	RO
18:16	<b>MaxVL</b> – Indicates the number of VLs supported. The Port supports VL <sub>0</sub> through VL <sub>MaxVL</sub> inclusive.	RO
18:16	<b>MaxVL</b> – Maximum number of VLs that the Port can support.	RO
23:19	<b>Reserved</b>	RO
30:24	<b>Maximum Time Slots</b> – Indicates the maximum number of time slots (minus one) that are supported when configured for time-based WRR VL Arbitration. For example, a value 000 0000b in this field indicates	RO

Bit Location	Register Description	Attributes
	the supported maximum number of time slots is 1 and a value of 111 1111b indicates the supported maximum number of time slot is 128.  This field is valid only when the VL Arbitration Capability field indicates that time-based WRR VL Arbitration is supported.	
31	<b>Reserved</b>	RO

#### 4.3.3.6. VL Arbitration Control (18h)

Table 4-46: Switch VL Arbitration Control

Bit Location	Register Description	Attributes
0	<p><b>Load VL Arbitration Table</b> – When Set, this bit updates the VL Arbitration logic from the VL Arbitration Table. This bit is valid only when the VL Arbitration Table is used by the selected VL Arbitration scheme (that is indicated by a Set bit in the VL Arbitration Capability field selected by VL Arbitration Select).</p> <p>Software sets this bit to signal hardware to update VL Arbitration logic with new values stored in VL Arbitration Table; clearing this bit has no effect. Software uses the VL Arbitration Table Status bit to confirm whether the new values of VL Arbitration Table are completely latched by the arbitration logic.</p> <p>This bit always returns 0b when read.</p> <p>Default value of this bit is 0b.</p>	RW
7:4	<b>Reserved</b>	RO
11:8	<p><b>VL Arbitration Select</b> – This field configures the Port to provide a particular VL Arbitration service.</p> <p>The permissible value of this field is a number corresponding to one of the asserted bits in the VL Arbitration Capability field.</p>	RO
15:12	<b>Reserved</b>	RO
23:16	<p><b>VL Strict Priority Arbitration</b> – This field contains one bit per VL. Bit 0 corresponds to VL0. Bit 7 corresponds to VL7.</p> <p>When a bit is Set, the corresponding VL is configured to arbitrate as Strict Priority based on VL number. When a bit is Clear, the corresponding VL is configured to arbitrate as normal priority (using the scheme selected by VL Arbitration Select).</p> <p>Among the VLs configured for strict priority, priority is based on increasing VL number. VL0 is the lowest strict priority, VL7 is the highest.</p> <p>Strict Priority VLs have priority over normal priority VLs.</p> <p>Behavior is Undefined if a VL configured for Strict Priority is also included in the VL Arbitration Table.</p> <p>If a VL is Disabled, the value of the corresponding bit in this field is ignored.</p> <p>Default value of this field is 0000 0000b.</p>	RW
31:24	<b>Reserved</b>	RO

#### 4.3.3.7. VL Arbitration Table Offset (1Ch)

Table 4-47: Switch VL Arbitration Table Offset

Bit Location	Register Description	Attributes
1:0	Reserved	RO
31:2	VL Arbitration Table Offset – DWORD Offset to the VL Arbitration Table	RO

#### 4.3.3.8. MR Error Status (20h)

Table 4-48: Switch MR Error Status

Bit Location	Register Description	Attributes
3:0	MR First Error Pointer	RO
4	MR Uncorrectable TLP Error Status	RW1C
5	MR Global Key Error Status	RW1C
14:6	Reserved	RsvdZ
15	MR DLLP Error Status – DLLP Errors are not logged	RW1C

#### 4.3.3.9. MR Error Control (22h)

Table 4-49: Switch MR Error Control

Bit Location	Register Description	Attributes
3:0	Reserved	RO
4	MR Uncorrectable TLP Error Mask	RW
5	MR Global Key Error Mask	RW
14:6	Reserved	RO
15	MR DLLP Error Mask	RW

#### 4.3.3.10. MR Error Log (24h to 34h)

These fields contain the TLP Prefix and TLP Header corresponding to the error described by the First Error Pointer in the MR Error Status register.

The value of these fields is undefined if the First Error Pointer is zero or points to a bit number that is not Set.

Headers are not logged and the First Error Pointer is not updated for DLLP Errors.



#### 4.3.3.11. PCI Bridge Control (N+00h)

These fields are the PCI Bridge controls that affect the Physical Port.

**Table 4-50: PCI Bridge Control**

Bit Location	Register Description	Attributes
0	<p><b>Secondary Bus Reset</b> – If the Link is a Downstream PCIe Link setting this bit causes the Port to initiate a PCI Express Hot Reset (TS1 with the Hot Reset bit Set). Clearing this bit causes the Port to remove the Hot Reset and attempt to bring the link back up. Flow Control is renegotiated after TS1 style Hot Reset is removed.</p> <p>If Bridge Control Physical is Set, this bit is identical to the Secondary Bus Reset bit in the PCI Bridge Control register.</p> <p>If the Link Direction Status does not equal 1 (i.e. Link is Upstream or Down) or if the Link Partner is MR bit is 1b, this bit must be 0b.</p>	RW
15:0	<b>Reserved</b>	RsvdP

#### 4.3.3.12. PCIe Capability Structure (N+02h)

These fields are the PCI Express controls that affect the Physical Port. Values in various Configuration Spaces are either Virtual values or map to these values (see the Bridge Controls Physical bit described in Section 4.3.6.2)

The layout of this structure is similar to the PCI Express Capability dropping the Root Capability, Control and Status words. All fields are implemented as defined in the *PCI Express Specification* except as indicated in [Table 4-51](#).

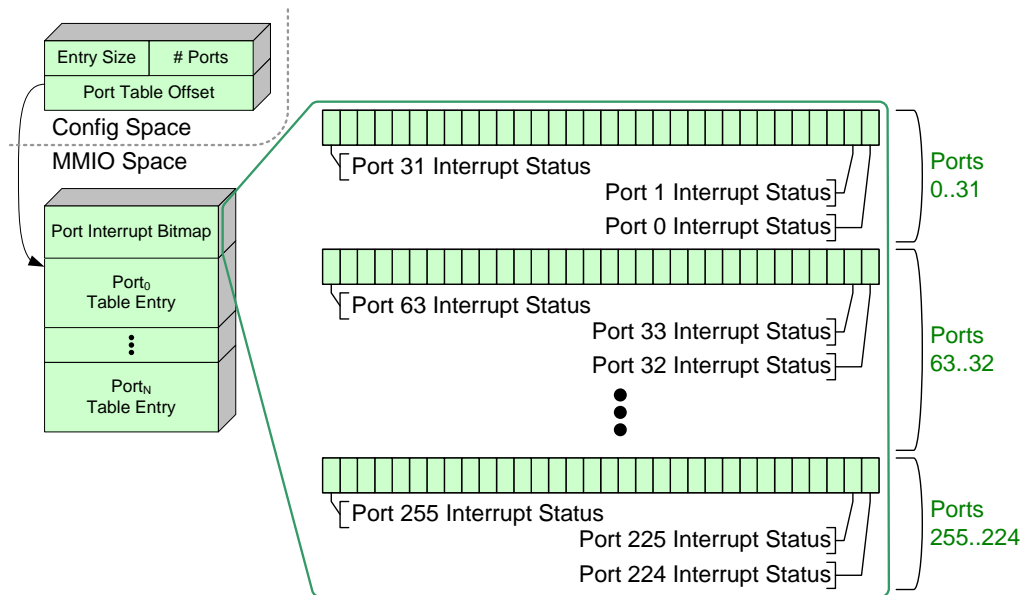
Deleted: Table 4-51

**Table 4-51: Port PCIe Capability Structure**

<b>Register</b>	<b>Field(s)</b>	<b>Attributes</b>
PCI Express Capabilities	Slot Implemented	Hwlnit. Reflects Physical Hot-Plug.
Device Capabilities	Phantom Functions Supported	Must be 00b
Device Capabilities	Captured Slot Power Limit Value Captured Slot Power Value	Must be 0h if Link Direction Status is non-zero (i.e. Link Down or Downstream Switch Port). If Base PCIe Link, reflects the value received. If MR Link reflects the value received on VH0, if VH0 is Upstream for this port, otherwise 0h.
Device Capabilities	Function Level Reset Capability	Must be 0b.
Device Control	Correctable Error Reporting Enable Non-Fatal Error Reporting Enable Fatal Error Reporting Enable Unsupported Request Reporting Enable	Not Implemented. Errors are raised and controlled within a VH.
Device Control	Enable Relaxed Ordering	Not Implemented. Controlled within a VH.
Device Control	Max_Payload_Size	Not Implemented. Controlled within a VH.
Device Control	Extended Tag Field Enable	Not Implemented. Controlled within a VH.
Device Control	Phantom Function Enable	Must be 0b.
Device Control	Enable No Snoop	Not Implemented, Controlled within a VH.
Device Control	Max_Read_Request_Size	Not Implemented. Controlled within a VH.
Device Status	Correctable Error Detected Non-Fatal Error Detected Fatal Error Detected Unsupported Request Detected Unsupported Request Detected	Not Implemented. Errors are raised and controlled within a VH.
Device Status	Transactions Pending	Not Implemented. Meaningful only within a VH.
Link Capabilities	Surprise Down Error Reporting Capable	Must be 1b.
Link Capabilities	Port Number	Not Implemented. In MR-IOV, Port Number is the index into the Port table.
Slot Capabilities	Attention Button Present Power Controller Present MRL Sensor Present Attention Indicator Present Power Indicator Present Hot-Plug Surprise Hot-Plug Capable Electromechanical Interlock Present	Hwlnit. Reflects Physical Hot-Plug capabilities
Slot Capabilities	Slot Power Limit Value	Must be 0h if Link Direction Control is 0 (i.e. Upstream Switch Port). If MR Link,

Register	Field(s)	Attributes
	Slot Power Limit Scale	Set_Slot_Power_Limit messages resulting from writing this value will be sent on VH0 if VH0 is Downstream for this port.
Slot Control	all fields	Reflect Physical Slot / Hot-Plug controls. Software Notification causes the Physical Hot-Plug Interrupt Pending bit in the Port Status register to be Set.
Device Capabilities 2	Completion Timeout Ranges Supported Completion Timeout Disable Supported ARI Forwarding Supported	Not Implemented. Implemented within a VH.
Device Control 2	Completion Timeout value Completion Timeout Disable ARI Forwarding Enable	Not Implemented. Implemented within a VH.

#### 4.3.3.13. Port Interrupt Status Bitmap (minus 20h)



Note: Port Interrupt Bitmap is always 32 bytes. This size supports the maximum size MR Switch. Port 0 is always at byte Physical\_Port\_Table\_Offset - 32. Bits for ports that are not present read as zero. Write 1 to Clear.

**Figure 4-8: Port Interrupt Status Bitmap**

The Port Interrupt Status bitmap precedes the Port Table. It is always 32 bytes (supporting switches with a maximum of 256 ports).

Bits in this table are Read Only. A bit is Set to indicate the Port has an interrupt pending and Clear otherwise. These Interrupt Status bits are cleared either by clearing the appropriate Port Interrupt Pending bit or by masking the interrupt using the Port Interrupt Enable.

An MSI Interrupt is requested on any zero to one transition of any of these bits.

Bits corresponding to Ports that are not Present are Read Only Zero.

#### 4.3.4. Switch VL Arbitration Table

Switch and Device VL Arbitration tables are identical. See Section 4.3.7 for details.

### 4.3.5. Switch VS Table

The VS Table contains up to 256 VS Table Entries. This table is located in memory starting at a location determined by the VS Table Offset field in the MR-IOV Capability.

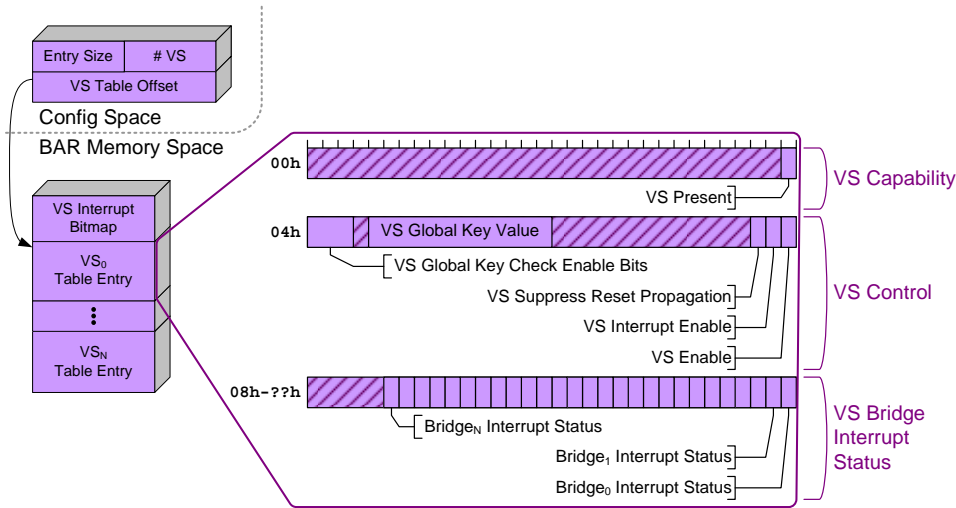


Figure 4-9: VS Table

#### 4.3.5.1. VS Capability and Status (00h)

Table 4-52: Switch VS Capability and Status

Bit Location	Register Description	Attributes
0	<b>VS Present</b> – Read only bit indicating that the VS is implemented. When VS Enable is zero, this bit may be controlled in a Vendor Specific mechanism to allow for flexible silicon implementation.	RO
31:1	<b>Reserved</b>	RO

#### 4.3.5.2. VS Control (04h)

Table 4-53: Switch VS Capability and Status

Bit Location	Register Description	Attributes
0	<b>VS Enable</b> – If set, indicates that MR-PCIM is using this VS. When set, the capabilities of the VS may not change. When clear, Vendor Specific mechanisms may change the capabilities offered by this VS. If the Switch does not support changing VS capabilities, this bit may be read only with the value one.	RW
1	<b>VS Interrupt Enable</b> – If Set, the VS Interrupt Summary bit corresponding to this VS will be affected when any of the Bridge N Interrupt Status bits in this VS are Set. If Clear, the Bridge N Interrupt Status bits have no effect on the corresponding VS Interrupt Summary bit.	RW
2	<b>VS Suppress Reset Propagation</b> – If Set, the automatic sending of Hot Reset or Reset DLLPs downstream is suppressed. If Clear, DL_Down, Hot Reset and/or Reset DLLPs received on the upstream Bridge of this VS cause downstream Bridges to send Hot Reset.  Suppressing Reset Propagation can be used to ensure that a failure in the management VH does not prematurely reset the entire MR Topology.  Suppressing Reset Propagation does not affect TLP discarding. TLPs destined to or from the upstream Bridge will be discarded as a result of the upstream Bridge being down or in Hot Reset. Forwarding of TLPs between downstream ports of the VS is not affected by the Reset state of the upstream Bridge.	RW
15:3	<b>Reserved</b>	RO
27:16	<b>VS Global Key Value</b> – Expected Global Key Value to for TLPs associated with the VS.  This value is inserted in TLPs originated by the VS.  This value is inserted in TLPs entering the VS from a non-MR Enabled Link.  If enabled, Global Keys for TLPs associated with the VS are checked against this value.	RW
28	<b>Reserved</b>	RO
31:29	<b>VS Global Key Check Enable</b> – Enables checking of the Global Key in the TLP Prefix against the Global Key Value for TLPs associated with the VS.  This is a three bit field that enables checks at various points in TLP processing.  Bit 29 enables the entering check. This validates forwarded TLPs as they are received by the Switch. This error is reported in the Port where the TLP entered the switch. Implementing this check is optional. If not implemented, this bit is read only zero.  Bit 30 enables the exiting check. This validates forwarded TLPs as they are transmitted by the Switch. This error is reported in the Port	RW

Bit Location	Register Description	Attributes
	<p>where the TLP exited the Switch. Implementing this check is optional. If not implemented, this bit is read only zero.</p> <p>Bit 31 enables the terminating check. This validates TLPs addressing the VS and TLPs being forwarded to a Base PCIe Link. This error is reported in the Port where the TLP entered the Switch. Every VS must implement this bit.</p> <p>A Switch may implement the bit 29 check, the bit 30 check, both checks or neither check. When implemented, a check must be implemented in every VS.</p>	

#### 4.3.5.3. VS Bridge Interrupt Status (08h to ??h)

This field contains one bit per P2P Bridge in the VS. Bit 0 of the first DWORD corresponds to VS Bridge Table Entry 0 (i.e. the Upstream Bridge). Bit 1 corresponds to the VS Bridge Table Entry 0. This field contains  $\text{INT}((\text{Num\_Bridge\_Table\_Entries} + 31) / 32)$  DWORDs.

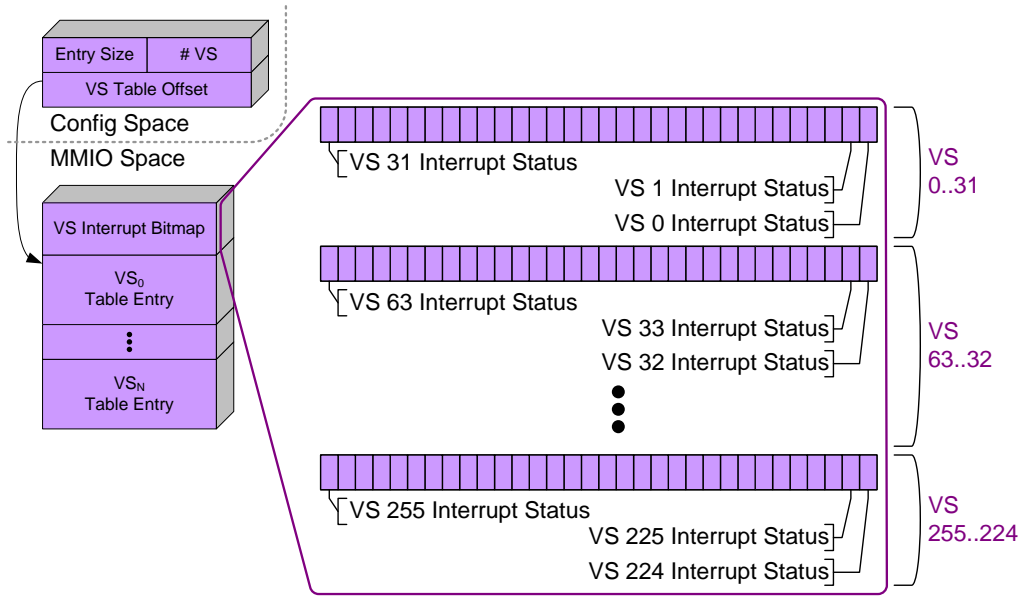
Bits in this field are Set only when the any of the following bits are Set in the VS Bridge Table.

- ☐ VS Status Changed
- ☐ Attention Indicator State Changed
- ☐ Power Indicator State Changed
- ☐ Power Controller State Changed

Bits in this field are Cleared by clearing the one of these “Changed” bits.

If VS Interrupt Enable is Set and any of the bits in this field are Set, the corresponding VS Interrupt Summary bitmap bit is Set.

#### 4.3.5.4. VS Interrupt Status Bitmap (minus 20h)



Note: VS Interrupt Bitmap is always 32 bytes. This size supports the maximum size MR Switch. VS 0 is always at byte VS\_Table\_Offset - 32. Bits for VSs that are not present are zero. The Bitmap is read only. An entry is 1 if a Global Key Fault is set or if any BridgeN Interrupt Status bit is set in the corresponding VS Table Entry.

**Figure 4-10: VS Interrupt Status Bitmap**

The VS Interrupt Status bitmap precedes the VS Table. It is always 32 bytes (supporting switches with a maximum of 256 Virtual Switches).

Bits in this table are Read Only. A bit is Set to indicate the VS has an interrupt pending and Clear otherwise. These Interrupt Status bits are cleared by clearing the VS BridgeN Interrupt bit Pending bit using the VS Bridge Table or by masking the interrupt using the VS Interrupt Enable.

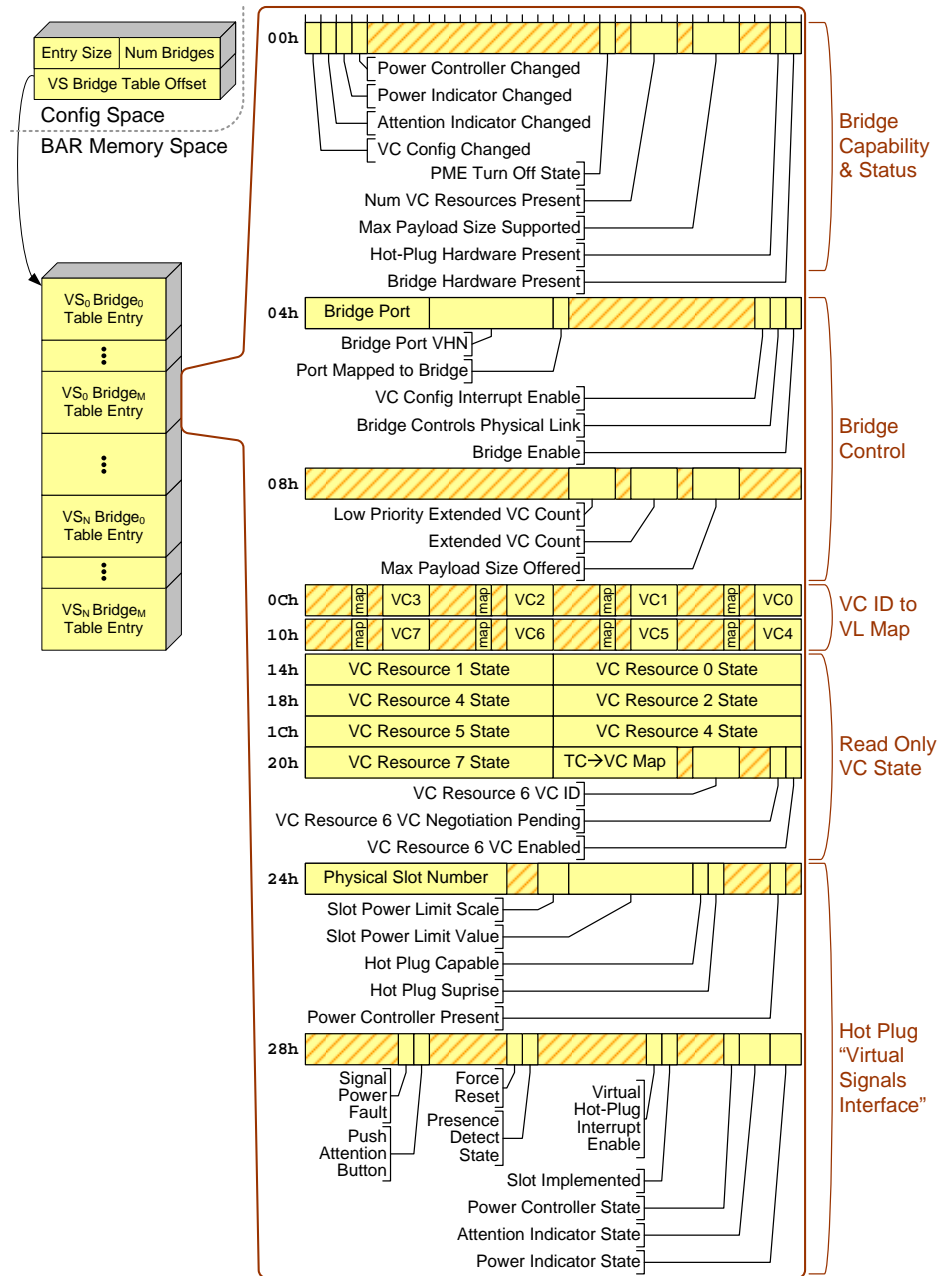
An MSI Interrupt is requested on any zero to one transition of any of these bits.

Bits corresponding to Virtual Switches that are not Present are Read Only Zero.

#### 4.3.6. Switch VS Bridge Table

The VS Bridge Table contains up to 256 Bridge Table Entries. This table is located in memory starting at a location determined by the VS Bridge Table Offset field in the MR-IOV Capability.





**Figure 4-11: VS Bridge Table**

VS Bridge Table Entry 0 corresponds to the upstream P2P Bridge of the VS. This bridge is always present.

VS Bridge Table Entry 1 corresponds to the downstream P2P Bridge located at Device 0, Function 0 on the VS internal bus. VS Bridge Table Entry  $N$  ( $1 \leq N \leq 32$ ) corresponds to the downstream P2P Bridge located at Device  $N-1$ , Function 0 on the VS Internal bus. VS Bridge Table Entries above 32 are used for P2P Bridges located at non-zero Function numbers as shown in the following table.

VS Bridge Table Entry $N$	Device	Function
1..32	$N-1$	0
33..64	$N-33$	1
65..96	$N-65$	2
97..128	$N-97$	3
129..160	$N-129$	4
161..192	$N-161$	5
193..224	$N-193$	6
225..256	$N-225$	7

#### 4.3.6.1. VS Bridge Capability and Status (00h)

Table 4-54: Switch VS Bridge Capability and Status

Bit Location	Register Description	Attributes
0	<p><b>Bridge Present</b> – Set if the hardware supports a bridge in this slot. The VS Bridge Table has the same number of Table Entries associated with each VS. This bit allows the table to be sparse and to be populated in a Vendor Specific manner. This bit is not allowed to change if VS Enabled is set.</p> <p>The Upstream Bridge is always present in any enabled VS.</p> <p>When VS Enable is clear, Vendor Specific mechanisms may change the value of this field.</p>	RO
1	<p><b>Hot-Plug Hardware Present</b> – This field indicates whether the P2P Bridge supports Hot-Plug. If clear, Hot-Plug hardware is not present and the remainder of the “signals interface” bits are read only zero. This bit is always Clear for the upstream Bridge of a VS.</p> <p>When VS Enable is clear, Vendor Specific mechanisms may change the value of this field.</p>	RO
3:2	<b>Reserved</b>	RO
6:4	<b>Max Payload Size Supported</b> – Returns the maximum Payload Size supported by the hardware.	RO
7	<b>Reserved</b>	RO
10:8	<p><b>Num VC Resources Hardware Present</b> – Indicates the index of the last VC Resource array structure implemented in the Type 1 Configuration Header associated with this Bridge in the VH. The value 0 indicates one VC Resource is provided. The value 7 indicates that all 8 VC Resources are provided.</p> <p>This value indicates the number that the hardware implements. MR-PCIM software may offer a lower number to the VH by setting Extended VC Count.</p> <p>When VS Enable is clear, Vendor Specific mechanisms may change the value of this field.</p>	RO
14:11	<b>Reserved</b>	RO
15	<p><b>PME Turn Off State</b> – Set when this bridge completes the PME Turn Off Handshake. Cleared when this bridge sends or receives any TLP other than PME_Turn_Off or PME_TO_Ack or when this bridge enters Reset.</p> <p>The PME Turn Off handshake completes when an Upstream bridge sends or a Downstream Bridge receives a PME_TO_Ack message.</p>	RO
27:16	<b>Reserved</b>	RO
28	<b>Power Controller State Changed</b> – This field indicates to MR-PCIM that the Power Controller State has been changed by software in the VH.	RW1C
29	<b>Power Indicator State Changed</b> – This field indicates to MR-PCIM	RW1C

Bit Location	Register Description	Attributes
	that the Power Indicator State has been changed by software in the VH.	
30	<b>Attention Indicator State Changed</b> – This field indicates to MR-PCIM that the Attention Indicator State has been changed by software in the VH.	RW1C
31	<b>VC Config Changed</b> – Indicates that software running in the VH changed the VC Configuration of some VC Resource associated with this Bridge.  This bit is Set if a VC Resource is enabled or disabled. This bit is also Set if the VC ID or TC to VC Map is changed while the VC is Enabled.	RW1C

#### 4.3.6.2. VS Bridge Control (04h and 08h)

Table 4-55: Switch VS Bridge Control 1

Bit Location	Register Description	Attributes
0	<p><b>Bridge Enable</b> – If set, the associated P2P Bridge is visible in the associated VS and the associated Type 1 Configuration Header is accessible.</p> <p>If clear, the associated P2P Bridge is disabled. Accesses by software in the VS to Memory and Configuration Space of the Bridge as well as Memory, Configuration and I/O Space of Functions below the Bridge will return UR.</p> <p>Clearing Bridge Enable for an Upstream VS Bridge Table entry blocks access to bridges below the upstream bridge and thus has the effect of hiding the entire VS from the view of software running in the VH.</p> <p>Enabling a VS is a two step process. First enable the VS. This ensures that Vendor Specific mechanisms will not alter the Bridge Present bits during the rest of this process. Then configure and enable specific bridges in the VS starting with downstream bridges and finishing with the upstream bridge.</p> <p>If Bridge Present is clear this field is Read Only Zero.</p> <p>The default value of this field is Vendor Specific.</p>	RW
1	<p><b>Bridge Controls Physical Link</b> – If Set and this Bridge is mapped to some Port's VH0, the fields in this Bridge's Type 1 Header control the Physical Link, and are not Virtual. Otherwise, the fields this Bridge's Type 1 Header are Virtual. The fields involved are all fields from the PCI Express Capability and selected additional fields from the Type 1 Header (see section 4.3.3.8 for the list of fields affected).</p> <p>If this bit is Set, the registers in the Type 1 Header are the same as the registers in the Switch Port Table Entry corresponding to the Port mapped to this Bridge. Changes to Type 1 Header bits for the bridge where this bit is Set affects the Type 1 Header and also affects the equivalent register in the Port Table. Similarly, changes in the Port Table registers are visible in the Type 1 Header corresponding to the Bridge where this bit is Set.</p> <p>By setting this bit, MR-PCIM is ceding control of the physical link to software running in the associated VH. This is necessary when a switch is being used as a Base PCIe switch. This may also be useful for Base PCIe links attached to MR Switches.</p> <p>This Bridge is mapped to VH0 of some Port if (1) the Port Mapped to Bridge bit is Set, (2) the Bridge VHN field is 0 and (3) the Bridge Port field contains a valid Port number.</p>	RW
2	<p><b>VC Config Interrupt Enable</b> – If Set, the VC Config Changed bit can trigger an interrupt. If Clear, VC Config Changed will not trigger an interrupt.</p>	
14:3	<b>Reserved</b>	RO
15	<p><b>Port Mapped to Bridge</b> – If Clear, no Port is mapped to this Bridge.</p>	RW

Bit Location	Register Description	Attributes
	<p>The Bridge VHN and Bridge Port fields are not used. Software operating in the VH sees the Data Link Layer Link Active bit Clear and link is in the virtual DL_Inactive state.</p> <p>If Set, a Port is mapped to this Bridge. The Bridge VHN and Bridge Port fields are used. The virtual Data Link Layer Link Active bit and the virtual link state track the physical link state of the mapped Port.</p> <p>Changes to the Virtual Data Link Layer Link Active bit, either through changes to this bit or changes to the physical link state cause a Data Link State Change event in the virtual Hot-Plug controller.</p> <p>The default value of this field is Vendor Specific. If Bridge Enable is Clear, this field is Read Only Zero.</p>	
23:16	<p><b>Bridge VHN</b> – This field indicates the Port VHN associated with this P2P Bridge. This value must be less or equal to than value of NumVH in the Port contained in Bridge Port. Hardware ignores the value of this field if the Port Mapped to Bridge bit is Clear. The default value of this field is Vendor Specific.</p>	RW
31:24	<p><b>Bridge Port</b> – This field indicates the Port associated with this P2P Bridge. This value must correspond to an enabled Port. Hardware ignores the value of this field if the Port Mapped to Bridge bit is Clear. The default value of this field is Vendor Specific.</p>	RW

Behavior is undefined if a Bridge Port / Bridge VHN combination is simultaneously mapped into more than one VS Bridge Table entry.

Table 4-56: Switch VS Bridge Control 2

Bit Location	Register Description	Attributes
3:0	<b>Reserved</b>	RO
6:4	<b>Max Payload Size Offered</b> – Indicates Maximum Payload Size offered in the Type 1 Header. This value must be less than or equal to the Max Payload Size Supported value.  Default value for this field is Vendor Specific.	RW
7	<b>Reserved</b>	RO
10:8	<b>Extended VC Count</b> – MR-PCIM must configure this value with the number of VC Resources that are offered to software operating in the VH. This value will be available as Extended VC Count in the VC Capability of the associated Type 1 Configuration Header.  This value must be less than or equal to Num VC Resources Hardware Present. Changing this value while Bridge Enable is set is undefined. The default value of this field is Vendor Specific.	RW
11	<b>Reserved</b>	RO
14:12	<b>Low Priority Extended VC Count</b> – MR-PCIM must configure this value with the value to be provided to software operating in the VH. This value has no hardware effect. This field's purpose is to inform software running in the VH of the relative priority of certain VCs.  This value must be less than or equal to the value set for Extended VC Count. Changing this value while Bridge Enable is set is undefined. The default value of this field is Vendor Specific.	RW
31:15	<b>Reserved</b>	RO

#### 4.3.6.3. VC ID to VL Map (0Ch and 10h)

These fields contain the Virtual Link to be used for traffic out this P2P Bridge for the indicated VC. VC to VL mapping is not needed and these fields are read only zero if the MaxVL value is zero for all Ports of the switch.

Software may not map multiple VCs to the same VL. Specifically, within a single VS Bridge Table entry, behavior is undefined if multiple enabled VL Map entries contain the same map value.

Table 4-57: Switch VS Bridge VC ID to VL Map 1

Bit Location	Register Description	Attributes
2:0	<b>VC0 VL Map</b> – Indicates the VL number used for VS traffic labeled VC0 transmitted via this VS Bridge. The default value of this field is Vendor Specific.	RW
3	<b>Reserved</b>	RO
4	<b>VC0 VL Map Enable</b> – Indicates that the VC0 VL Map field contains a valid VL number. The default value of this field is Vendor Specific. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
7:5	<b>Reserved</b>	RO
10:8	<b>VC1 VL Map</b> – Indicates the VL number used by VS traffic labeled VC1 transmitted via this VS Bridge. This field is Read Only Zero if Num VC Resources Hardware Present is 0. The default value of this field is Vendor Specific.	RW
11	<b>Reserved</b>	RO
12	<b>VC1 VL Map Enable</b> – Indicates that the VC1 VL Map field contains a valid VL number. This field is Read Only Zero if Num VC Resources Hardware Present is 0. The default value of this field is Vendor Specific. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
15:13	<b>Reserved</b>	RO
18:16	<b>VC2 VL Map</b> – Indicates the VL number used by VS traffic labeled VC2 transmitted via this VS Bridge. This field is Read Only Zero if Num VC Resources Hardware Present is 0 or 1. The default value of this field is Vendor Specific.	RW
19	<b>Reserved</b>	RO
20	<b>VC2 VL Map Enable</b> – Indicates that the VC2 VL Map field contains a valid VL number. This field is Read Only Zero if Num VC Resources Hardware Present is 0 or 1. The default value of this field is Vendor Specific. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
23:21	<b>Reserved</b>	RO
26:24	<b>VC3 VL Map</b> – Indicates the VL number used by VS traffic labeled VC3 transmitted via this VS Bridge. This field is Read Only Zero if	RW



Bit Location	Register Description	Attributes
	Num VC Resources Hardware Present is less than or equal to 2. The default value of this field is Vendor Specific.	
27	<b>Reserved</b>	RO
28	<b>VC3 VL Map Enable</b> – Indicates that the VC3 VL Map field contains a valid VL number. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 2. The default value of this field is Vendor Specific.  Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
31:29	<b>Reserved</b>	RO

Table 4-58: Switch VS Bridge VC ID to VL Map 2

Bit Location	Register Description	Attributes
2:0	<b>VC4 VL Map</b> – Indicates the VL number used by VS traffic labeled VC4 transmitted via this VS Bridge. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 3. The default value of this field is Vendor Specific.	RW
3	<b>Reserved</b>	RO
4	<b>VC4 VL Map Enable</b> – Indicates that the VC4 VL Map field contains a valid VL number. The default value of this field is Vendor Specific.  Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
7:5	<b>Reserved</b>	RO
10:8	<b>VC5 VL Map</b> – Indicates the VL number used by VS traffic labeled VC5 transmitted via this VS Bridge. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 4. The default value of this field is Vendor Specific.	RW
11	<b>Reserved</b>	RO
12	<b>VC5 VL Map Enable</b> – Indicates that the VC5 VL Map field contains a valid VL number. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 4. The default value of this field is Vendor Specific.  Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
15:13	<b>Reserved</b>	RO
18:16	<b>VC6 VL Map</b> – Indicates the VL number used by VS traffic labeled VC6 transmitted via this VS Bridge. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 5. The default value of this field is Vendor Specific.	RW
19	<b>Reserved</b>	RO
20	<b>VC6 VL Map Enable</b> – Indicates that the VC6 VL Map field contains a valid VL number. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 5. The default value of this field is Vendor Specific.	RW

Bit Location	Register Description	Attributes
	Hardware behavior on the 1b to 0b transition of this bit is undefined.	
23:21	<b>Reserved</b>	RO
26:24	<b>VC7 VL Map</b> – Indicates the VL number used by VS traffic labeled VC7 transmitted via this VS Bridge. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 6. The default value of this field is Vendor Specific.	RW
27	<b>Reserved</b>	RO
28	<b>VC7 VL Map Enable</b> – Indicates that the VC7 VL Map field contains a valid VL number. This field is Read Only Zero if Num VC Resources Hardware Present is less than or equal to 6. The default value of this field is Vendor Specific. Hardware behavior on the 1b to 0b transition of this bit is undefined.	RW
31:29	<b>Reserved</b>	RO

#### 4.3.6.4. VC Resource Fields (14h to 20h)

These fields return data from the VC Capability of the associated Type 1 Configuration Header. They allow MR-PCIM software to track the enabling and mapping of VCs with each VH.

VC Resource fields for resource numbers above Num VC Resource Hardware Present are not implemented and return 0 when read. VC Resource fields for resource numbers above Extended VC Count are Undefined.

VC Resource State 0 is located at offset 0Ch; VC Resource 1 is located at offset 0Eh; etc. Fields within VC Resource State are described in [Table 4-59](#).

**Deleted:** Table 4-59

Table 4-59: VC Resource State

Bit Location	Register Description	Attributes
0	<b>VC Enabled</b> – This field tracks the VC Enabled bit set by software operating in the VH. Per the PCI Express specification, VC Resource 0 is always enabled and thus VC Enabled for VC Resource 0 is always set.	RO
1	<b>VC Negotiation Pending</b> – This field tracks the VC Negotiation Pending bit view in the VH. This bit is Set when VC Enabled is Set and either no VL has been mapped to this VC (associated VL Map Enable bit is Clear) or Flow Control negotiation has not completed on the mapped VH and VL.	RO
3:1	<b>Reserved</b>	RO
6:4	<b>VC ID</b> – This field tracks the VC ID field set by software operating in the VH. Per the PCI Express specification, VC ID for VC Resource 0 is always 0.	RO
7	<b>Reserved</b>	RO
15:8	<b>TC to VC Map</b> – This field tracks the TC to VC Map field set by software operating in the VH. Per the PCI Express specification, bit 0 of this field is fixed and the remaining bits may be set by software.	RO

#### 4.3.6.5. Hot-Plug Virtual Signals Interface (24h and 28h)

These bits form the “Virtual Signals Interface” of the Virtual Hot-Plug controller. These registers allow MR-PCIM to indicate to software what Hot-Plug features are supported and to control those features.

See Chapter 6 Hot Plug for additional details.

Virtual Hot-Plug controller hardware is optional. Presence of hardware is indicated by the Hot Plug Hardware Present bit. If not hardware is present, the Slot Implemented bit is Read Only zero and some of the bits in this section are Undefined. Virtual Hot-Plug hardware is only present in downstream Ports.

If the Bridge Controls Physical Link bit is Set, the Virtual Hot-Plug Signals Interface Registers are not used and their content is Undefined.

Table 4-60: Virtual Hot-Plug Signals Interface 1

Bit Location	Register Description	Attributes
0	<b>Reserved</b>	RO
1	<p><b>Virtual Power Controller Present</b> – This field indicates to software in the VH that the virtual slot has a Power Controller. This is visible to software in the VH through the Power Controller Present field of the PCI Express Capabilities.</p> <p>If Set, a Virtual Power Controller exists in the VH. When software operating in the VH turns off power to the virtual slot using the Virtual Hot-Plug Power Controller, a per-VH Reset is automatically triggered to clean out state in downstream hardware.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RW
4:2	<b>Reserved</b>	RO
5	<p><b>Virtual Hot Plug Surprise</b> – This field indicates to software in the VH that the slot supports Surprise Hot-Plug events. Writing this field changes the Slot Capabilities field of the PCI Express Capabilities. The default value of this field is Vendor Specific.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RW
6	<p><b>Virtual Hot Plug Capable</b> – This field indicates to software in the VH that the slot is Hot-Plug Capable. Writing this field changes the Slot Capabilities field of the PCI Express Capabilities. The default value of this field is Vendor Specific.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RW
14:7	<p><b>Virtual Slot Power Limit Value</b> – This field contains the value written to the PCI Express Capability Slot Power Limit Value register by software in the VH. There is no other effect.</p> <p>If Hot-Plug Hardware Present is 0b or if Virtual Slot Implemented is 0b, this field is undefined.</p>	RO
16:15	<p><b>Virtual Slot Power Limit Scale</b> – This field contains the value written to the PCI Express Capability Slot Power Limit Scale register by software in the VH. There is no other effect.</p> <p>If Hot-Plug Hardware Present is 0b or if Virtual Slot Implemented is 0b, this field is undefined.</p>	RO
18:17	<b>Reserved</b>	RO
31:19	<p><b>Virtual Slot Number</b> – This field indicates the chassis slot number. It is echoed to software in the VH via the PCI Express Capabilities field. The default value of this field is Vendor Specific.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RW

Table 4-61: Hot-Plug Signals Interface 2

Bit Location	Register Description	Attributes
1:0	<p><b>Virtual Power Indicator State</b> – This field indicates the state of the Power Indicator as set by software in the VH. This corresponds to the Power Indicator Control field of the PCI Express Capability Slot Control Register.</p> <p>Changes to this value set the Power Indicator Changed bit.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RO
3:2	<p><b>Virtual Attention Indicator State</b> – This field indicates the state of the Attention Indicator as set by software in the VH. This corresponds to the Attention Indicator Control field of PCI Express Capability Slot Control Register.</p> <p>Changes to this value set the Attention Indicator Changed bit.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RO
4	<p><b>Virtual Power Controller State</b> – This field indicates the state of the Power Controller as set by software in the VH. This corresponds to the Power Controller Control bit of the Slot Control field of the PCI Express Capabilities.</p> <p>Changes to this value set the Power Controller Changed bit.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RO
7:5	<b>Reserved</b>	RO
8	<p><b>Virtual Slot Implemented</b> – This field indicates to software in the VH whether the P2P Bridge contains Hot-Plug support. Writing this field changes the Slot Implemented bit in the PCI Express Capabilities field.</p> <p>If Hot-Plug Hardware Present is 0b, this field is Read Only Zero.</p>	RW
9	<p><b>Hot Plug Signals Interrupt Enable</b> – If Set, the Power Controller Changed, Power Indicator Changed and Attention Indicator Changed bits can trigger an interrupt. If Clear, these fields will not trigger an interrupt.</p> <p>If Hot-Plug Hardware Present is 0b, this field is Read Only Zero.</p>	RW
16:10	<b>Reserved</b>	RO
17	<p><b>Virtual Presence Detect State</b> – This field indicates that the virtual slot has a card in it. Defined Encodings are:</p> <p>0b Virtual Slot Empty 1b Card Present in Virtual Slot</p> <p>The value of this field affects both the Presence Detect State and the Presence Detect Changed fields in the PCI Express Capability Slot Status Register.</p> <p>The Presence Detect State field contains the same value as this field.</p> <p>The Presence Detect Changed field is Set whenever this field changes state.</p> <p>If Hot-Plug Hardware Present is 0b, this field is undefined.</p>	RW

Bit Location	Register Description	Attributes
18	<p><b>Virtual Force Reset</b> – For downstream ports, when this field is Set, the Switch causes the Port / Port VHN mapped to this Bridge to enter Reset. The effect is the same as if software running the VH set the Secondary Bus Reset bit in the associated Type 1 Header.</p> <p>For upstream Ports, this field is Read Only Zero.</p> <p>For downstream Ports, this field is always present, even if Hot-Plug Hardware Present is 0b.</p>	RW
23:18	<b>Reserved</b>	RO
19	<p><b>Push Virtual Attention Button</b> – Writing 1 to this field Sets the Attention Button Pressed bit in the PCI Express Capability Slot Status Register. This simulates the press of the virtual Attention Button in the Virtual Hot-Plug Controller.</p> <p>Writing 0 to this field has no effect. If Hot-Plug Hardware Present is 0b, writing to this field has no effect. This field read as Zero.</p>	Write 1 to Trigger, Read Zero
20	<p><b>Signal Virtual Power Fault</b> – Writing 1 to this field Sets the Power Fault Detected bit in the PCI Express Capability Slot Status Register. This simulates a Power Fault condition in the Virtual Hot-Plug Controller.</p> <p>Writing 0 to this field has no effect. If Hot-Plug Hardware Present is 0b, writing to this field has no effect. This field reads as Zero.</p>	Write 1 to Trigger, Read Zero
31:21	<b>Reserved</b>	RO

#### 4.3.7. Misc. Switch Configuration Space Requirements

##### 4.3.7.1. ARI Support

Alternate RID Interpretation (ARI) support must be provided in all downstream PCI-to-PCI bridges of MRA Switches. Specifically, the ARI Forwarding Supported bit located in the Device Capabilities 2 register must be set and the ARI Forwarding Enable bit located in the Device Control 2 register must be implemented.

##### 4.3.7.2. BIST (switch)

MR-IOV Switches shall not support BIST.

Implementation Note: In a virtualized environment, BIST would also have to be virtualized. Since BIST is rarely used and not completely specified it was decided to remove it from MR-IOV.

## 4.4. VL Arbitration Table

The VL Arbitration Table is optional. It is identical in structure to the VC arbitration table in PCI Express.

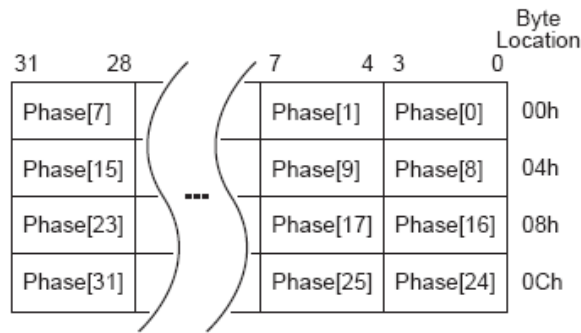
Switches and Devices use the same VL Arbitration Table structure. Switches and Devices differ in the location of the fields used to locate the VL Arbitration table and to configure the arbitration scheme used. For a switch, fields in the Port Table are used for this purpose. For a Device, fields in the MR-IOV Capability are used instead.

The VL Arbitration Table is a read-write register array that is used to store the arbitration table for VL Arbitration. This register array is valid for all Functions when the selected VL Arbitration uses a WRR table. If it exists, the VL Arbitration Table is located by the VL Arbitration Table Offset field.

The VL Arbitration Table is a register array with fixed-size entries of 4 bits. [Figure 4-12](#) depicts the table structure of an example VL Arbitration Table with 32 phases. Each 4-bit table entry corresponds to a phase within a WRR arbitration period. The definition of table entry is depicted in [Table 4-62](#). The lower 3 bits (bits 0-2) contain the VL value, indicating that the corresponding phase within the WRR arbitration period is assigned to the Virtual Channel indicated by the VL (must be a valid VL that corresponds to an enabled VL).

The highest bit (bit 3) of the table entry is reserved. The length of the table depends on the selected VL Arbitration as shown in [Table 4-63](#).

When the VL Arbitration Table is used by the default VL Arbitration method, the default values of the table entries must be all zero to ensure forward progress for the default VL (with VL of 0).



OM14489

Figure 4-12: Example VL Arbitration Table with 32 Phases

Table 4-62: Definition of the 4-bit Entries in the VL Arbitration Table

Bit Location	Register Description	Attributes
2:0	VL	RW
3	Reserved	RsvdP

Table 4-63: Length of the VL Arbitration Table

VC Arbitration Select	VC Arbitration Table Length (in # of Entries)
001b	32
010b	64
011b	128

## 4.5. Performance Monitoring and Statistics Collection

This section describes the registers and tables used to control the optional MR-IOV Performance Monitoring and Statistics Collection Capability.

Switch and Device usage is identical except where noted.

An overview of the registers and tables associated with this capability is shown in [Figure 4-13](#). Detailed descriptions are provided in the following sections. The functional behavior is described in Section 8.3.

Deleted: Figure 4-13



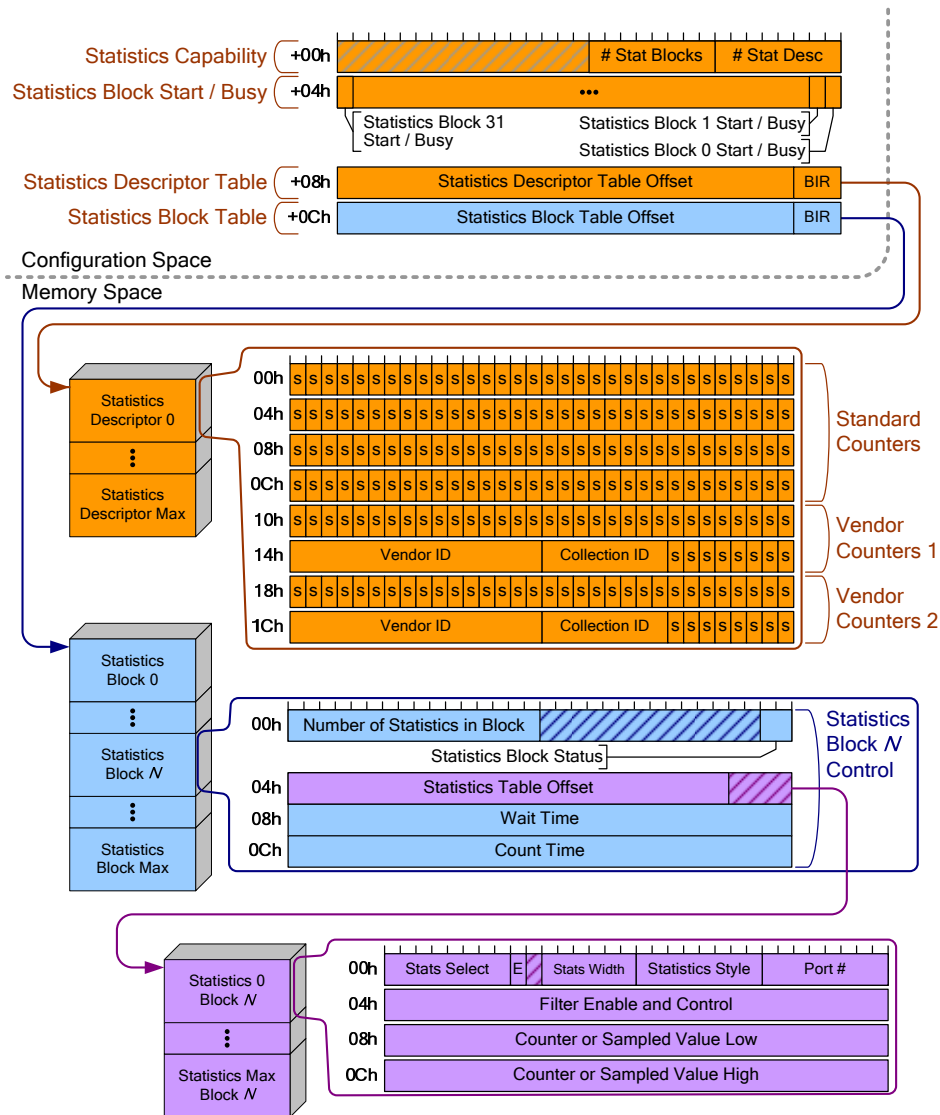


Figure 4-13: Performance Monitoring and Statistics Collection Tables

#### 4.5.1. Configuration Space Fields

The following fields are located in the Switch and Device MR-IOV Capabilities. Starting offsets of these fields within these two capabilities are different.

In addition to the fields described below, each Capability contains Statistics Interrupt Status and Statistics Interrupt Enable bits.

#### 4.5.1.1. *Statistics Capability (+00h)*

These fields define the sizes of the Statistics Tables pointed to by the MR-IOV Capability.

If the Performance Monitoring and Statistics Collection Capability is not implemented, these fields are Read Only Zero.

**Table 4-64: Statistics Table Sizes**

Bit Location	Register Description	Attributes
7:0	<b>Number of Statistics Descriptors</b> – Indicates the number of entries in the Statistics Descriptor Table.  Statistics support is optional for all components but strongly encouraged for MRA Switches. If not supported, this field is Zero.	RO
15:8	<b>Number of Statistics Blocks</b> – Indicates the number of entries in the Statistics Block Table. This value must be less than or equal to 32.  Statistics support is optional for all components but strongly encouraged for MRA Switched. If not supported, this field is Zero.	RO
31:16	<b>Reserved</b>	RO

#### 4.5.1.2. Statistics Block Start / Busy (+04h)

These fields contain a bit for each supported Statistics Block.

**Table 4-65: Statistics Start / Busy**

Bit Location	Register Description	Attributes
0	<p><b>Statistics Block 0 Start / Busy</b> – If idle, writes of 1b initiate the statistics collection processing for Statistics Block 0. The behavior of initiating statistics collection for a Statistics Block that is not idle is undefined.</p> <p>If not idle, writes of 0b terminate the statistics collection process for Statistics Block 0. Termination of the statistics collection process is not immediate; therefore, following termination of the statistics collection process, this field should be read to confirm that termination has completed and that the Statistics Block is idle.</p> <p>When read, indicates the busy status of the associated Statistics Block. The value 1b indicates the Statistics Block is busy (i.e. either waiting or counting). The value 0b indicates the Statistics Block is idle.</p> <p>If Number of Statistics Blocks is zero, this field is Read Only Zero.</p>	RW
1	<p><b>Statistics Block 1 Start / Busy</b> – Controls Statistics Block 1. If Number of Statistics Block is 0 or 1, this field is Read Only Zero.</p>	RW
2	<p><b>Statistics Block 2 Start / Busy</b> – Controls Statistics Block 2. If Number of Statistics Block is less than or equal to 2, this field is Read Only Zero.</p>	RW
...	...	...
31	<p><b>Statistics Block 31 Start / Busy</b> – Controls Statistics Block 31. If Number of Statistics Block is less than or equal to 31, this field is Read Only Zero.</p>	RW

4.5.1.3. *Statistics Descriptor Table Offset (+08h)*

**Table 4-66: Statistics Descriptor Table Offset**

Bit Location	Register Description	Attributes																												
2:0	<p><b>Statistics Descriptor Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's Statistics Descriptor Table into Memory Space.</p> <p><b>BIR Value Base Address register</b></p> <table><tr><td>0</td><td>BAR0</td><td>10h</td><td></td></tr><tr><td>1</td><td>BAR1</td><td>14h</td><td></td></tr><tr><td>2</td><td>BAR2</td><td>18h</td><td>(Device Only)</td></tr><tr><td>3</td><td>BAR3</td><td>1Ch</td><td>(Device Only)</td></tr><tr><td>4</td><td>BAR4</td><td>20h</td><td>(Device Only)</td></tr><tr><td>5</td><td>BAR5</td><td>24h</td><td>(Device Only)</td></tr><tr><td>6..7</td><td>Reserved</td><td></td><td></td></tr></table> <p>For a 64-bit Base Address register, the BIR indicates the lower DWORD.</p> <p>For Switch usage, the values 2..5 are Reserved as well.</p>	0	BAR0	10h		1	BAR1	14h		2	BAR2	18h	(Device Only)	3	BAR3	1Ch	(Device Only)	4	BAR4	20h	(Device Only)	5	BAR5	24h	(Device Only)	6..7	Reserved			RO
0	BAR0	10h																												
1	BAR1	14h																												
2	BAR2	18h	(Device Only)																											
3	BAR3	1Ch	(Device Only)																											
4	BAR4	20h	(Device Only)																											
5	BAR5	24h	(Device Only)																											
6..7	Reserved																													
31:3	<p><b>Statistics Descriptor Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the Statistics Descriptor Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.</p>	RO																												

#### 4.5.1.4. Statistics Block Table Offset (+0Ch)

Table 4-67: Statistics Block Table Offset

Bit Location	Register Description	Attributes																												
2:0	<p><b>Statistics Block Table BIR</b> – Indicates which one of a function's Base Address registers, located beginning at 10h in Configuration Space, is used to map the Function's Statistics Block Table into Memory Space.</p> <p><b>BIR Value Base Address register</b></p> <table><tr><td>0</td><td>BAR0</td><td>10h</td><td></td></tr><tr><td>1</td><td>BAR1</td><td>14h</td><td></td></tr><tr><td>2</td><td>BAR2</td><td>18h</td><td>(Device Only)</td></tr><tr><td>3</td><td>BAR3</td><td>1Ch</td><td>(Device Only)</td></tr><tr><td>4</td><td>BAR4</td><td>20h</td><td>(Device Only)</td></tr><tr><td>5</td><td>BAR5</td><td>24h</td><td>(Device Only)</td></tr><tr><td>6..7</td><td>Reserved</td><td></td><td></td></tr></table> <p>For a 64-bit Base Address register, the BIR indicates the lower DWORD.</p> <p>For Switch usage, the values 2..5 are Reserved as well.</p>	0	BAR0	10h		1	BAR1	14h		2	BAR2	18h	(Device Only)	3	BAR3	1Ch	(Device Only)	4	BAR4	20h	(Device Only)	5	BAR5	24h	(Device Only)	6..7	Reserved			RO
0	BAR0	10h																												
1	BAR1	14h																												
2	BAR2	18h	(Device Only)																											
3	BAR3	1Ch	(Device Only)																											
4	BAR4	20h	(Device Only)																											
5	BAR5	24h	(Device Only)																											
6..7	Reserved																													
31:3	<p><b>Statistics Block Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the Statistics Block Table. The lower 3 BIR bits are masked off (set to zero) by software to form a 32-bit QWORD-aligned offset.</p>	RO																												

#### 4.5.2. Statistics Descriptor Table

The Statistics Descriptor Table describes sets of statistics that may be recorded by Statistics Counters.

This table contains up to 256 entries. Each entry is 256 bits. The entire Statistics Descriptor Table is Read Only and constant.

A Descriptor Table Entry describes generate statistics recording capabilities. Each Statistics Counter contains the index of a Descriptor Table Entry that indicates the statistics recording capabilities associated with that counter.

Each Descriptor Table entry contains 208 Supported bits (S bits). If an S bit is 1b, then Statistics Counters that point to that Descriptor Table entry may be configured to record the statistic associated with the S bit. This configuration is done by setting the Statistics Select field of the counter to the bit number of the S bit.

S bits 127:0 represent statistics defined in this specification. S bits 167:128 and 231:192 represent Vendor Specific statistics.

Table 4-68: Statistics Descriptor Table Entry

Bit Location	Register Description	Attributes
127:0	<b>Standard S Bits</b> – Indicates S bits whose meaning is defined by this specification.	RO
167:128	<b>Group 1 Vendors Specific S Bits</b> – Indicates S bits whose meaning is defined by the Vendor listed in Group 1 Vendor ID further qualified by Group 1 Collection ID. If no Group 1 S bits are supported, this field is hardwired to zero.	RO
175:168	<b>Group 1 Collection ID</b> – Vendor Specific identifier that defines the meaning of bits 167:128. If no Group 1 S bits are supported, this field is hardwired to zero.	RO
191:176	<b>Group 1 Vendor ID</b> – Indicates the Vendor that defined the meaning of bits 175:128. If no Group 1 S bits are supported, this field is hardwired to zero.	RO
231:192	<b>Group 2 Vendors Specific S Bits</b> – Indicates S bits whose meaning is defined by the Vendor listed in Group 2 Vendor ID further qualified by Group 2 Collection ID. If no Group 2 S bits are supported, this field is hardwired to zero.	RO
239:232	<b>Group 2 Collection ID</b> – Vendor Specific identifier that defines the meaning of bits 231:192. If no Group 2 S bits are supported, this field is hardwired to zero.	RO
255:240	<b>Group 2 Vendor ID</b> – Indicates the Vendor that defined the meaning of bits 239:192. If no Group 2 S bits are supported, this field is hardwired to zero.	RO

The following nomenclature is used to describe counters:

- ❑ Standard statistics are designated CSEL[*n*] where *n* is in the range [0..127] (inclusive).
- ❑ Vendor Specific statistics are designated CSEL[Vendor ID, Collection ID, *n*]. Vendor ID is assigned by the PCI SIG. Collection ID a Vendor defined value used to select a set of S bit definitions. The value *n* is in the range [0..39] (inclusive) and the corresponding S bit number is *n* + 168 (if mapped using Group 1) or *n* + 192 (if mapped using Group 2). The meaning of a Vendor Specific statistic is not affected by whether it is mapped using Group 1 or Group 2.

This mechanism allows a single counter to support any mixture of standard events and vendor defined events from up to two sets of S bit definitions (from either the same or different Vendors).

#### 4.5.2.1. Standard Statistics

**Table 4-69** Error! Reference source not found. contains Standard Statistics defined by this standard. Items marked Sample correspond to sampled values while items marked Count correspond to counted values. Standard filters are defined in Section 4.5.2.2.

Deleted: Table 4-69

**Table 4-69: Standard Statistics**

S Bit Number	Description	Count / Sample	Applicable Filter(s)
0	<b>Transmitted TLPs</b> – Counts non-nullified TLPs transmitted by the Port. <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	<b>Optional TLP Filters:</b> VH, VL, TLP Type
1	<b>Transmitted TLP DWORDs</b> – Counts DWORDs of non-nullified TLPs transmitted by the Port. This includes framing symbols and all bytes sent (i.e. STP to END inclusive). <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	<b>Optional TLP Filters:</b> VH, VL, TLP Type
2	<b>Transmitted IDLE Symbols</b> – Counts the number of IDLE symbols transmitted by the Port. <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	None
3	<b>Transmitted DLLPs</b> – Counts the number of DLLPs sent by this Port <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	<b>Optional DLLP Filters:</b> DLLP Type
5	<b>Any TLP Blocked by VL</b> – Number of Symbol times a TLP is blocked from transmission due to lack of VL flow control credits	Count	<b>Optional Credit Filters:</b> TLP Type <b>Required Credit Filters:</b> VL
6	<b>Any TLP Blocked by {VH, VL}</b> – Number of Symbol times a TLP is blocked from transmission due to the lack of {VH, VL} flow control credits.	Count	<b>Optional Credit Filters:</b> TLP Type <b>Required Credit Filters:</b> VH, VL
7	<b>Any TLP Blocked by VL or {VH, VL}</b> – Number of Symbol times a TLP is blocked from transmission due either to the lack of VL flow control credits or the lack of {VH, VL} flow control credits.	Count	<b>Optional Credit Filters:</b> TLP Type <b>Required Credit Filters:</b> VH, VL
9	<b>All TLP Blocked by VL</b> – Number of Symbol times all TLPs are blocked from transmission due to lack of VL flow control credits (i.e., the lack of VL flow control credits results in no TLP being transmitted on the wire).	Count	<b>Optional Credit Filters:</b> TLP Type <b>Required Credit Filters:</b> VL
10	<b>All TLP Blocked by {VH, VL}</b> – Number of Symbol	Count	<b>Optional Credit Filters:</b>

S Bit Number	Description	Count / Sample	Applicable Filter(s)
	times all TLPs are blocked from transmission due to the lack of {VH, VL} flow control credits (i.e., the lack of {VH, VL} flow control credits results in no TLP being transmitted on the wire).		TLP Type <b>Required Credit Filters:</b> VH, VL
11	<b>All TLP Blocked by VL or {VH, VL}</b> – Number of Symbol times all TLPs are blocked from transmission due either to the lack of VL flow control credits or the lack of {VH, VL} flow control credits (i.e., the lack of VL or {VH, VL} flow control credits results in no TLP being transmitted on the wire).	Count	<b>Optional Credit Filters:</b> TLP Type <b>Required Credit Filters:</b> VH, VL
32	<b>Available VL Transmit Credits</b> – Number of available transmit credits associated with a VL computed as: $(\text{CREDIT\_LIMIT} - \text{CREDITS\_CONSUMED}) \bmod 2^{\text{[Field Size]}}$	Sample	<b>Required Credit Filters:</b> VL, Credit Type
33	<b>Available {VH, VL} Transmit Credits</b> – Number of available transmit credits associated with a {VH, VL} computed as: $(\text{CREDIT\_LIMIT} - \text{CREDITS\_CONSUMED}) \bmod 2^{\text{[Field Size]}}$	Sample	<b>Required Credit Filters:</b> VL, VH, Credit Type
64	<b>Received TLPs</b> – Counts non-nullified TLPs received by this port Port. <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	<b>Optional TLP Filters:</b> VH, VL, TLP Type
65	<b>Received TLP DWORDs</b> – Counts DWORDs of non-nullified TLPs received by this Port. This includes framing symbols and all bytes received (i.e. STP to END inclusive). <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	<b>Optional TLP Filters:</b> VH, VL, TLP Type
66	<b>Received Idle Symbols</b> – Counts the number of IDLE symbols received by this Port. <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	None
67	<b>Received DLLPs</b> – Counts the number of DLLPs received from by this Port. <i>Required to be implemented by at least two Statistics Counters per port.</i>	Count	<b>Optional DLLP Filters:</b> DLLP Type
96	<b>Available VL Receive Credits</b> – Number of available receive credits associated with a VL computed as: $(\text{CREDITS\_ALLOCATED} - \text{CREDITS\_RECEIVED}) \bmod 2^{\text{[Field Size]}}$	Sample	<b>Required Credit Filters:</b> VL, Credit Type
97	<b>Available {VH, VL} Receive Credits</b> – Number of available receive credits associated with a {VH, VL} computed as: $(\text{CREDITS\_ALLOCATED} - \text{CREDITS\_RECEIVED}) \bmod 2^{\text{[Field Size]}}$	Sample	<b>Required Credit Filters:</b> VL, VH, Credit Type



#### 4.5.2.2. Standard Filters

TLP Filtering consists of three filters: TLP Type, VL and VH.

**Table 4-70: TLP Filters**

Bits	Filter	Description
0	TLP Type	Completion – If Set, Completion TLPs are <i>not</i> included (i.e., filtered out) If TLP Type filtering is not supported, this field is hardwired to zero.
1	TLP Type	Non-Posted – If Set, Non-Posted TLPs are <i>not</i> included (i.e., filtered out). If TLP Type filtering is not supported, this field is hardwired to zero.
2	TLP Type	Posted – If Set, Posted TLPs are <i>not</i> included (i.e., filtered out). If TLP Type filtering is not supported, this field is hardwired to zero.
15:3		Reserved
23:16	VH	VH Value – If VH filtering is enabled, this field contains the VH to include (i.e., filtered in). If VH filtering is not enabled, this field is ignored and all VHs are included. If VH filtering is not supported, this field is hardwired to zero.
26:24	VL	VL Value – IF VL filtering is enabled, this field contains the VL to include (i.e., filtered in). If VL filtering is not enabled, this field is ignored and all VLs are included. If VL filtering is not supported, this field is hardwired to zero.
29:27		Reserved
30	VH	VH Filter Enable – If Set, VH filtering is enabled. If Cleared, VH filtering is not enabled and all VHs are included. If VH filtering is not supported, this field is hardwired to zero.
31	VL	VL Filter Enable – If Set, VL filtering is enabled. If Cleared, VL filtering is not enabled and all VLs are included. If VL filtering is not supported, this field is hardwired to zero.

Credit Filtering consists of three filters: Credit Type, VL and VH. Unsupported credit filters are hardwired to zero.

**Table 4-71: Credit Filters**

Bits	Filter	Description
0	Credit Type	Completion Header – If Set, Completion Header Credits are <i>not</i> included (i.e., filtered out).
1	Credit Type	Non-Posted Header – If Set, Non-Posted Header Credits are <i>not</i> included (i.e., filtered out).
2	Credit Type	Posted Header – If Set, Posted Header Credits are <i>not</i> included (i.e., filtered out).
3	Credit Type	Completion Data – If Set, Completion Data Credits are <i>not</i> included (i.e., filtered out).
4	Credit Type	Non-Posted Data – If Set, Non-Posted Data Credits are <i>not</i> included (i.e., filtered out).
5	Credit Type	Posted Data – If Set, Posted Data Credits are <i>not</i> included (i.e., filtered out).
15:6		Reserved
23:16	VH	VH Value – Contains the VH to include (i.e. filtered in).
26:24	VL	VL Value – Contains the VL to include (i.e., filtered in).
31:27		Reserved

DLLP Filtering is optional and consists of a number of filters. Unsupported DLLP filters are hardwired to zero.

**Table 4-72: DLLP Filters**

Bits	Filter	Description
0	DLLP Type	Ack – If Set, Ack DLLPs are <i>not</i> included (i.e., filtered out).
1	DLLP Type	Nak – If Set, Nak DLLPs are <i>not</i> included (i.e., filtered out).
2	DLLP Type	Reset – If Set, Reset DLLPs are <i>not</i> included (i.e., filtered out).
3	DLLP Type	MRInit – If Set, MRInit DLLPs are <i>not</i> included (i.e., filtered out).
4	DLLP Type	Flow Control Initialization – If Set, the following DLLPs are <i>not</i> included (i.e., filtered out): InitFC1 InitFC2 MRInitFC1_VL MRInitFC1_VH MRInitFC2_VL MRInitFC2_VH
5	DLLP Type	Flow Control Update – If Set, UpdateFC and MRUpdateFC DLLPs are <i>not</i> included (i.e., filtered out).
6	DLLP Type	ASPM L1 – If Set, PM_Active_State_Request_L1 and PM_Request_Ack DLLPs are <i>not</i> included (i.e., filtered out).
7	DLLP Type	PM L1 L23 – If Set, PM_Enter_L1 and PM_Enter_L23 DLLPs are <i>not</i> included (i.e., filtered out).
8	DLLP Type	Vendor Specific – If Set, vendor Specific DLLPs are <i>not</i> included (i.e., filtered out).
31:9		Reserved

### 4.5.3. Statistics Block Table

This table contains an entry for each supported Statistics Block. Up to 32 Statistics Blocks may be supported by a component.

#### 4.5.3.1. Statistics Block Capability (00h)

These fields describe the capabilities of the associated Statistics Block.

Table 4-73: Statistics Block Capability

Bit Location	Register Description	Attributes								
1:0	<p><b>Statistics Block Status</b> – Indicates whether the statistics block is busy and if so, whether the block is waiting or counting. Values in this field correspond to the states of the statistics collection process.</p> <p><b>Values are:</b></p> <table><tr><td>00b</td><td>Idle</td></tr><tr><td>10b</td><td>Waiting</td></tr><tr><td>11b</td><td>Counting</td></tr><tr><td>01b</td><td>Reserved</td></tr></table> <p>Note that the upper bit of this field matches the value read from the Statistics Block Start / Busy register.</p>	00b	Idle	10b	Waiting	11b	Counting	01b	Reserved	RO
00b	Idle									
10b	Waiting									
11b	Counting									
01b	Reserved									
15:2	<b>Reserved</b>	RO								
31:16	<b>Statistics Table Size</b> – Indicates the number of entries contained in this Statistics Table associated with this Statistics Block.	RO								

#### 4.5.3.2. Statistics Table Offset (04h)

Table 4-74: Statistics Table Offset

Bit Location	Register Description	Attributes
3:0	<b>Reserved</b>	RO
31:4	<b>Statistics Table Offset</b> – Used as an offset from the address contained by one of the Function's Base Address registers to point to the base of the Statistics Table. The Base Address register used is selected by the Statistics Block BIR located in the MR-IOV Capability.	RO

#### 4.5.3.3. Statistics Wait Time (08h)

Table 4-75: Statistics Wait Time

Bit Location	Register Description	Attributes
15:0	<p><b>Waiting Period</b> – Indicates the time, in microseconds, of the waiting period. The waiting period is defined as the time from statistics collection initiation to the start of the counting period.</p> <p>A value of zero indicates no waiting period.</p>	RW
31:16	<b>Reserved</b>	RO

#### 4.5.3.4. Statistics Count Time (0Ch)

Table 4-76: Statistics Count Time

Bit Location	Register Description	Attributes
23:0	<b>Counting Period</b> – Indicates the time, in microseconds, of the counting period. The counting period is defined as the time during which selected events are counted and is the time from the end of the waiting period to the idle period.  Note: A value of all ones (i.e. FFFFFFFh) is interpreted as infinite. An infinite counting period ends when requested by software (i.e. the corresponding Statistics Block Start bit is Cleared).  A value of zero corresponds to no counting period (this is useful for sampled values).	RW
31:24	<b>Reserved</b>	RO

#### 4.5.4. Statistics Counter Table

Associated with each Statistics Block Table Entry is a Statistics Counter Table. The Statistics Counter Table contains one entry for each implemented Statistics Counter associated with a Statistics Block. The registers described in this section form a Statistics Counter Table entry.

#### 4.5.4.1. Statistics Capability and Control (00h)

Table 4-77: Statistics Capability and Control

Bit Location	Register Description	Attributes
7:0	<b>Port Number</b> – Indicates which port the counter is associated with. For a Device, this field must be zero. For a switch, this value contains an index into the Port Table.	RO
15:8	<b>Statistics Descriptor Index</b> – Contains the index of the Statistics Descriptor Table entry that describes the statistics recording capabilities of this counter.	RO
21:16	<b>Counter Width</b> – Indicates the width of the counter. Value is counter width-1 (i.e. a 32 bit counter contains 31 in this field).  If the counter supports any standard statistics, the implemented counter width must be 32 bits or greater. If the counter supports only Vendor Specific statistics, the counter width can be any value.	RO
22	<b>Reserved</b>	RO
23	<b>Counter Enable</b> – When Set, the counter is enabled. When Cleared, the counter is disabled and certain fields (defined below) are undefined.  Software should disable unused counters to reduce power consumption.  This bit is allowed to be hardwired to 1b if the counter is always enabled.  The default value of this field is Vendor Specific.	RW
31:24	<b>Statistics Select</b> –Determines what statistic software has selected for this counter to record. This field contains the bit number of one of the Supported bits of the Statistics Descriptor entry selected by Statistics Descriptor Index.  The counter value is undefined if the value of this field indicates an unsupported counter (i.e. the value does not correspond to an S bit or the associated S bit is 0b).  Bits in this field may be Read Only if they are not needed. For example, a Statistics Descriptor Index that supports S bits 16, 17 and 18 need only implement this field as 000100WWb where W represents read/write bits. Following this rule to its extreme, if a Statistics Descriptor Index supports exactly one S bit, this entire field may be Read Only.  The default value of this field is Vendor Specific. This field is undefined when the counter is disabled.	RW

#### 4.5.4.2. Statistics Filter Enable and Control (04h)

Table 4-78: Statistics Filter Enable and Control

Bit Location	Register Description	Attributes
31:0	<b>Filter Enable and Control</b> – The meaning of this field depends on the Statistics Style and Statistics Select fields. The description of each counter defines the meaning of the filters associated with it. This field is undefined when the counter is disabled. The default value of this field is Vendor Specific.	RW

#### 4.5.4.3. Statistics Counter Low (08h)

Table 4-79: Statistics Counter Low

Bit Location	Register Description	Attributes
Width:0	<b>Count Value Low</b> – Indicates the lower 32 bits of the counter value. Unused bits, as indicated by Counter Width, are Read Only Zero. The counter value is undefined when the counter is disabled or Busy (i.e., in the waiting or counting period). The default value of this field is Vendor Specific.	RO
31:Width	<b>Reserved</b>	RO

#### 4.5.4.4. Statistics Counter High (0Ch)

Table 4-80: Statistics Counter High

Bit Location	Register Description	Attributes
Width-32:0	<b>Count Value High</b> – Indicates the upper bits of the counter value. Unused bits, as indicated by Counter Width, are Read Only Zero. The counter value is undefined when the counter is disabled or Busy (i.e., in the waiting or counting period). The default value of this field is Vendor Specific.	RO
31:Width-32+1	<b>Reserved</b>	RO



## 5. Error Handling

### 5.1. PCIe Error mapping to MR

The basic rules for error detection, logging and reporting are unchanged from PCIe. The only change is which VH(s) should be affected.

**Table 5-1: Physical Layer Error List**

Error Name	Error Type	Detecting Agent Action (PCIe)	Detecting Agent Action (MR IOV)
Receiver Error	Correctable	<i>Receiver (if checking):</i> Send ERR_COR to Root Complex.	<i>Receiver:</i> Same as PCIe but send on all enabled VHs not in Reset.

**Table 5-2: Data Link Layer Error List**

Error Name	Error Type	Detecting Agent Action (PCIe)	Detecting Agent Action (MR IOV)
Bad TLP	Correctable	<i>Receiver:</i> Send ERR_COR to Root Complex.	<i>Receiver:</i> Same as PCIe but send to all enabled VHs not in Reset.
Bad DLLP	Correctable	<i>Receiver:</i> Send ERR_COR to Root Complex.	<i>Receiver:</i> Same as PCIe but send on all enabled VHs not in Reset.
Replay Timeout	Correctable	<i>Transmitter:</i> Send ERR_COR to Root Complex.	<i>Transmitter:</i> Same as PCIe but send on all enabled VHs not in Reset
REPLAY NUM Rollover	Correctable	<i>Transmitter:</i> Send ERR_COR to Root Complex.	<i>Transmitter:</i> Same as PCIe but send on all enabled VHs not in Reset
Data Link Layer Protocol Error	Uncorrectable (Fatal)	If checking, send ERR_FATAL to Root Complex.	Same as PCIe but send to all enabled VHs not in Reset



Table 5-3: Transaction Layer Error List

Error Name	Error Type	Detecting Agent Action (PCIe)	Detecting Agent Action (MV IOV)
Poisoned TLP Received	Uncorrectable (Non-Fatal)	<p><i>Receiver:</i></p> <p>Send ERR_NONFATAL to Root Complex, or ERR_COR for the Advisory Non-Fatal Error cases</p> <p>Log the header of the Poisoned TLP.</p>	<p><i>Receiver:</i></p> <p>Same as PCIe.</p> <p>Error message sent only for affected VH.</p> <p>Log the header for the affected VH only.</p>
ECRC Check Failed	Uncorrectable (Non-Fatal)	<p><i>Receiver (if ECRC checking supported):</i></p> <p>Send ERR_NONFATAL to Root Complex, or ERR_COR for the Advisory Non-Fatal Error case</p> <p>Log the header of the TLP that encounter the ECRC error.</p>	<p><i>Receiver (if ECRC checking supported):</i></p> <p>Same as PCIe</p> <p>Error message sent only for affected VH.</p> <p>Log the header for the affected VH only.</p>
Unsupported Request (UR)	Uncorrectable (Non-Fatal)	<p><i>Request Receiver:</i></p> <p>Send ERR_NONFATAL to Root Complex, or ERR_COR for the Advisory Non-Fatal Error case</p> <p>Log the header of the TLP that caused the error.</p>	<p><i>Request Receiver:</i></p> <p>Same as PCIe</p> <p>Error message sent only for affected VH.</p> <p>Log the header for the affected VH only.</p>
Completion Timeout	Uncorrectable (Non-Fatal)	<p><i>Requester:</i></p> <p>Send ERR_NONFATAL to Root Complex, or ERR_COR for the Advisory Non-Fatal Error case</p>	<p><i>Requester:</i></p> <p>Same as PCIe</p> <p>Error message sent only for affected VH.</p>
Completer Abort	Uncorrectable (Non-Fatal)	<p><i>Completer:</i></p> <p>Send ERR_NONFATAL to Root Complex, or ERR_COR for the Advisory Non-Fatal Error case</p> <p>Log the header of the Request that encountered the error.</p>	<p><i>Completer:</i></p> <p>Same as PCIe</p> <p>Error message sent only for affected VH.</p> <p>Log the header for the affected VH only.</p>
Unexpected Completion	Uncorrectable (Non-Fatal)	<p><i>Receiver:</i></p> <p>Send ERR_COR to Root Complex. This is an Advisory Non-Fatal Error</p> <p>Log the header of the Completion that encountered</p>	<p><i>Receiver:</i></p> <p>Same as PCIe.</p> <p>Error message sent only for affected VH.</p> <p>Log the header for the affected VH only.</p>

Error Name	Error Type	Detecting Agent Action (PCIe)	Detecting Agent Action (MV IOV)
		the error.	
Receiver Overflow	Uncorrectable (Fatal)	<i>Receiver (if checking):</i> Send ERR_FATAL to Root Complex.	<i>Receiver (if checking):</i> Same as PCIe. Send ERR_FATAL to all Root Complex that have a VC on the affected link mapped to the affected VL.
Flow Control Protocol Error	Uncorrectable (Fatal)	<i>Receiver (if checking):</i> Send ERR_FATAL to Root Complex.	<i>Receiver (if checking):</i> Same as PCIe. Send ERR_FATAL to all Root Complex that have a VC on the affected link mapped to the affected VL.
Malformed TLP	Uncorrectable (Fatal)	<i>Receiver:</i> Send ERR_FATAL to Root Complex.  Log the header of the TLP that encountered the error.	<i>Receiver:</i> Same as PCIe.  Error message sent only for affected VH.  Log the header for the affected VH only.

## 5.2. MR Errors

Table 5-4: MR Error List

Error Name	Error Type	Detecting Agent Action (PCIe)	Detecting Agent Action (MR IOV)
Invalid TLP Prefix	Uncorrectable	N/A	<i>Receiver:</i> Signal MR Uncorrectable TLP Error, discard TLP, do not update flow control (VL / VH can't be trusted).
TLP received on VH in reset (sender of TLP has Acked entering Reset)	Uncorrectable	N/A	<i>Receiver:</i> Signal MR Uncorrectable TLP Error, discard TLP, update flow control normally.
TLP Prefix with Global Key Mismatch: TLP at destination or forwarded on PCIe Link	Uncorrectable	N/A	<i>Receiver:</i> Signal MR Global Key Error, discard TLP, update flow control normally
TLP Prefix with Global Key Mismatch: TLP being forwarded on MR Link	Correctable	N/A	<i>Receiver:</i> Signal MR Global Key Error, forward TLP normally.
TLP Prefix {VH, VL} that is invalid, not enabled or has not finished Flow Control Initialization	Uncorrectable	N/A	<i>Receiver:</i> Signal MR Uncorrectable TLP Error, discard TLP, do not update flow control (VL / VH can't be trusted).
MRUpdateFC for {VH, VL} that is invalid, not enabled or has not finished Flow Control Initialization	Correctable	N/A	<i>Receiver:</i> Signal MR DLLP Error, discard DLLP.
Invalid VH Group in Reset DLLP	Correctable	N/A	<i>Receiver:</i> Signal MR DLLP Error, discard DLLP.
Out of range Assert bit set in Reset DLLP	Correctable	N/A	<i>Receiver:</i> Signal MR DLLP Error, ignore the offending Assert bit(s) and process remainder

Error Name	Error Type	Detecting Agent Action (PCIe)	Detecting Agent Action (MR IOV)
			of Reset DLLP normally.



## 6. Hot Plug

### 6.1. MRA Switch

The Virtual Signals Interface registers in the VS Bridge Table provide the MR-PCIM end of the “Virtual Hot-Plug Signals Interface”. This section describes the connection between those fields and the PCI Express Slot registers of the corresponding Type 1 Header.

If the Bridge Controls Physical Link field in the VS Bridge Table is Clear, the VS Type 1 Header, slot registers are virtual and connect to the “Virtual Hot-Plug Signals Interface” described in Section 4.3.6.5.

If the Bridge Controls Physical Link field in the VB Bridge Table is Set, the VS Type 1 Header fields are physical and the external signals of the switch (if implemented). Specifically, in this mode, the Type 1 Header slot registers and the slot registers in the Port Table are identical. Changes to one register affects the other, external events are visible in both, there is a single set of “changed” bits and write 1 to clear to either register clears then.

The Type 1 header associated with each virtual downstream Switch Port shall contain the PCIe Slot Capabilities, Control and Status registers. These registers work exactly as in base PCIe as described in sections 7.8.9, 7.8.10 and 7.8.11.

For each Type 1 Header register, the following sections describe how the various bits interact with the “Virtual Hot-Plug Signals Interface”. Figures are extracted for reference from the *PCI Express Base Specification*.

Changes to the Virtual Hot-Plug Controller occur immediately. If Bridge Controls Physical Link is Clear, the No Command Complete Support bit in the Type 1 Header is always Set.

### 6.1.1. PCI Express Capability: Slot Capability Register

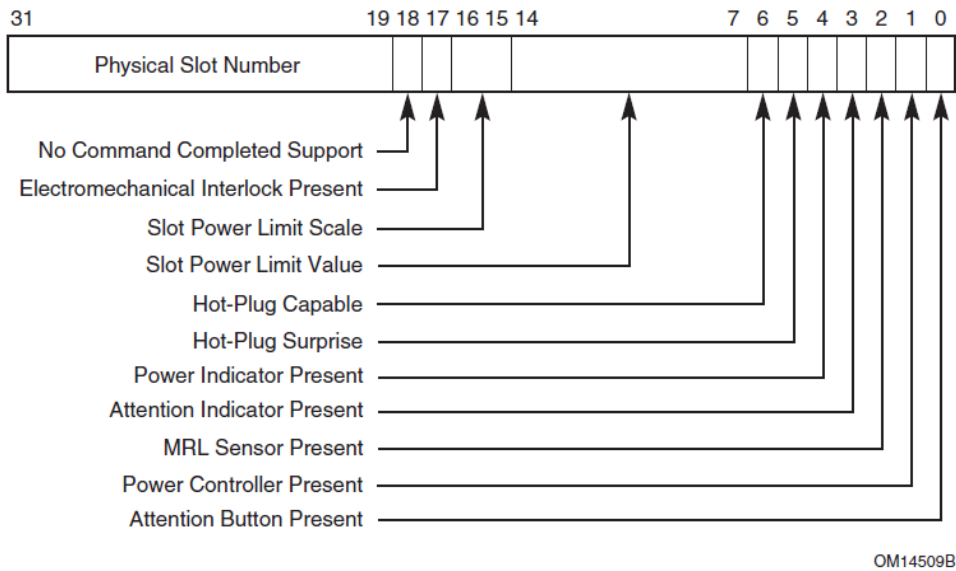


Figure 6-1: Slot Capabilities Register (PCIe Figure 7-18)

Table 6-1: Virtual Mapping: PCIe Slot Capabilities Register

Bits	Field Name	If Virtual (Bridge Controls Physical Link = 0b)	If Physical (Bridge Controls Physical Link = 1b)
0	Attention Button Present		Base (HWInit)
1	Power Controller Present	Bit 1 Signals Interface 1, Power Controller Present	Base (HWInit)
2	MRL Sensor Present	0b	Base (HWInit)
3	Attention Indicator Present	1b	Base (HWInit)
4	Power Indicator Present	1b	Base (HWInit)
5	Hot-Plug Surprise	Bit 5 Signals Interface 1, Virtual Hot Plug Surprise	Base (HWInit)
6	Hot Plug Capable	Bit 6 Signals Interface 1, Virtual Hot Plug Capable	Base (HWInit)
14:7	Slot Power Limit Value	Bits 14:7 Signals Interface 1, Virtual Slot	Base (HWInit)

Bits	Field Name	If Virtual (Bridge Controls Physical Link = 0b)	If Physical (Bridge Controls Physical Link = 1b)
		Power Limit Value	
16:15	Slot Power Limit Scale	Bits 16:15 Signals Interface 1, Virtual Slot Power Limit Scale	Base (HWInit)
17	Electromechanical Interlock Present	0b	Base (HWInit)
18	No Command Completed Support	0b	Base (HWInit)
31:19	Physical Slot Number	Bits 31:19 Signals Interface 1, Virtual Slot Number	Base (HWInit)

## 6.1.2. PCI Express Capability: Slot Control Register

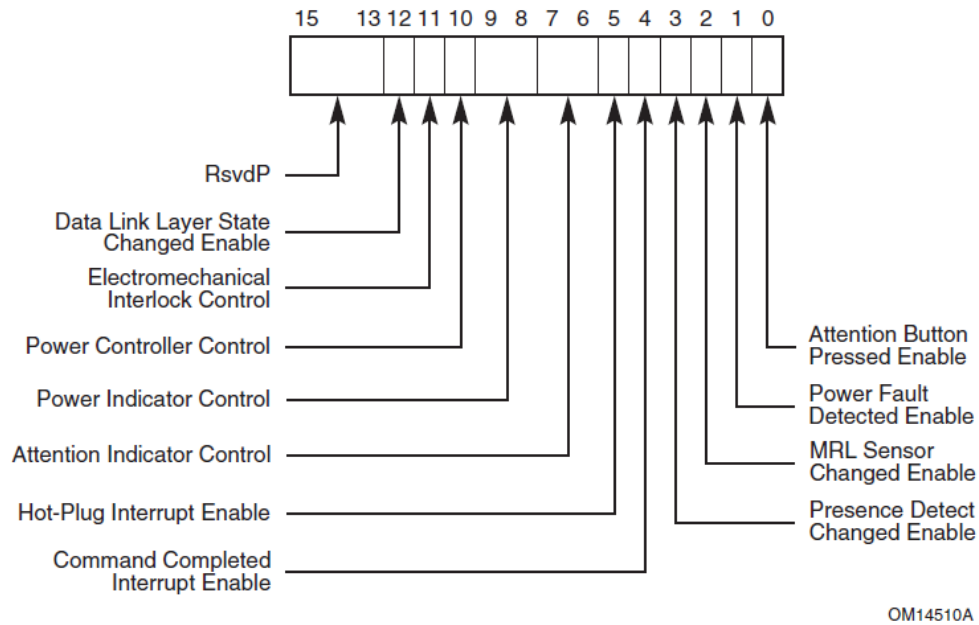


Figure 6-2: Slot Control Register (PCIe Figure 7-19)

Table 6-2: Virtual Mapping: PCIe Slot Control Register

Bits	Field Name	If Virtual (Bridge Controls Physical Link = 0b)	If Physical (Bridge Controls Physical Link = 1b)
0	Attention Button Pressed Enable	Implement	Base
1	Power Fault Detected Enable	Implement	Base
2	MRL Sensor Changed Enable	0b	Base
3	Presence Detect Changed Enable	Implement	Base
4	Command Completed Interrupt Enable	0b	Base
5	Hot-Plug Interrupt Enable <b>Note:</b> This bit controls Interrupts delivered in	Implement	Base



Bits	Field Name	If Virtual (Bridge Controls Physical Link = 0b)	If Physical (Bridge Controls Physical Link = 1b)
	the VH. It is unrelated to the similarly named Hot-Plug Signals Interrupt Enable bit. The latter governs interrupts delivered to MR-PCIM in the Management VS.		
7:6	Attention Indicator Control	Bits 3:2, Signals Interface 2, Virtual Attention Indicator State	Base
9:8	Power Indicator Control	Bits 1:0, Signals Interface 2, Virtual Power Indicator State	Base
10	Power Controller Control	Bit 4 Signals Interface 2, Virtual Power Controller State  Turning off Virtual Power State also causes the associated link to see a VH Reset (this is the same effect as setting Secondary Bus Reset).	Base
11	Electromechanical Interlock Control	0b	Base
12	Data Link Layer Changed Enable	Implement	Base
15:13	Reserved	0b	Base

### 6.1.3. PCI Express Capability: Slot Status Register

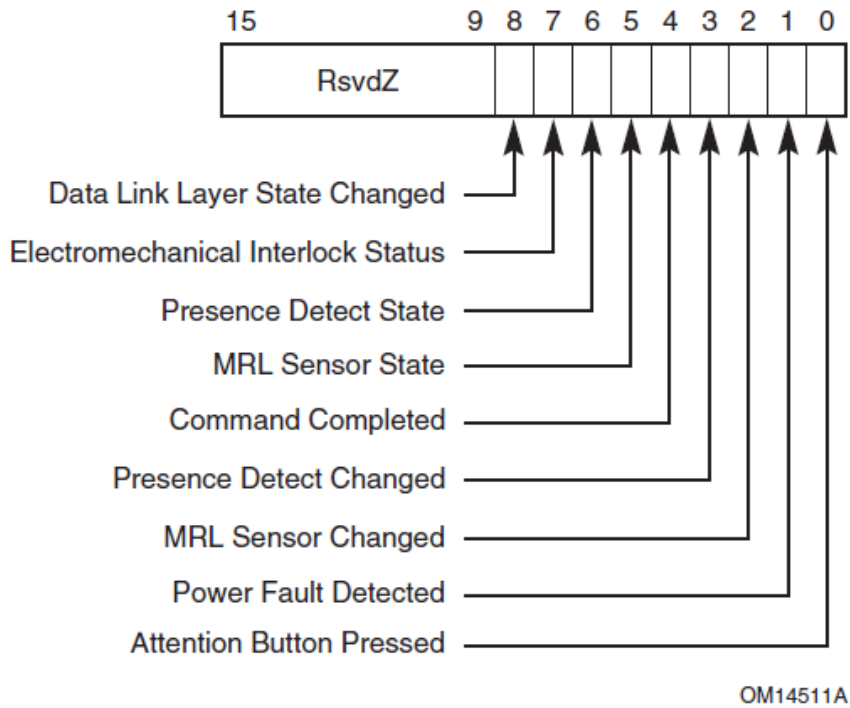


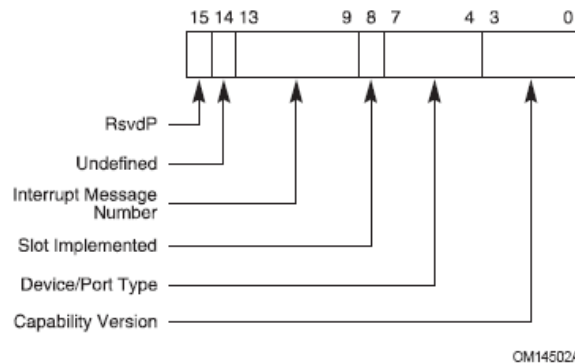
Figure 6-3: Slot Status Register (PCIe Figure 7-20)

Table 6-3: Virtual Mapping: PCIe Slot Control Register

Bits	Field Name	If Virtual (Bridge Controls Physical Link = 0b)	If Physical (Bridge Controls Physical Link = 1b)
0	Attention Button Pressed	Set if 1b written to Bit 24, Signals Interface 2, Push Virtual Attention Button	Base
1	Power Fault Detected	Set if 1b written to Bit 25 Signals Interface 2, Signal Virtual Power Fault	Base
2	MRL Sensor Changed	0b	Base
3	Presence Detect Changed	Set on transition of Bit 17 Signals Interface 2, Virtual Presence Detect	Base

Bits	Field Name	If Virtual (Bridge Controls Physical Link = 0b)	If Physical (Bridge Controls Physical Link = 1b)
		State	
4	Command Completed	0b	Base
5	MRL Sensor State	0b	Base
6	Presence Detect State	Bit 17, Signals Interface 2, Virtual Presence Detect State	Base
7	Electromechanical Interlock Status	0b	Base
8	Data Link Layer State Changed	Set on transition of Bit 16 Signals Interface 2, Virtual Data Link State	Base
15:9	Reserved	0b	Base

#### 6.1.4. PCI Express Capability: Device Capabilities Register



**Figure 6-4: PCI Express Capabilities Register (PCIe Figure 7-11)**

**Table 6-4: Virtual Mapping: PCIe Capabilities Register**

Bits	Field Name	If Virtual (Bridge Controls Physical Link = 0b)	If Physical (Bridge Controls Physical Link = 1b)
3:0	Capability Version	Base	Base
7:4	Device / Port Type	Base	Base
8	Slot Implemented	Bit 8, Signals Interface 2, Virtual Slot Implemented	Base
13:9	Interrupt Message Number	Base	Base
14	Undefined	Base	Base
15	Reserved	Base	Base

#### 6.1.5. Virtual Hot-Plug Signals Interface Registers

Each virtual downstream Switch Port has a Virtual Signals Interface as defined in Section 4.3.6.5. Registers controlling this hardware are located in the VS Bridge Table. They provide the interface between MR-PCIM Hot-Plug software and Hot-Plug software running in the VH. This interface allows MR-PCIM to:

1. Provide configuration and status information the virtual slot registers

2. Push virtual buttons of the Virtual Hot-Plug controller (i.e. change bits in the Virtual Slot Status register)
3. Detect Indicator and Control changes made to the Virtual Hot-Plug controller (i.e. detect certain changes to the Virtual Slot Control and Virtual Slot Capabilities registers)

The following register fields are defined for this function:

Bit Field	MR-PCIM View	Purpose
Virtual Slot Number	Read / Write	1
Virtual Slot Power Limit Scale / Value	Read Only	3
Virtual Slot Implemented	Read / Write	1
Virtual Hot-Plug Capable	Read / Write	1
Virtual Hot-Plug Surprise	Read / Write	1
Virtual Data Link State	Read / Write	2
Virtual Power Controller State	Read Only	3
Virtual Power Controller State Changed	Read / Write One to Clear	3
Virtual Power Controller Present	Read / Write	1
Virtual Power Indicator State	Read Only	3
Virtual Power Indicator State Changed	Read / Write One to Clear	3
Virtual Attention Indicator State	Read Only	3
Virtual Attention Indicator State Changed	Read / Write One to Clear	3
Virtual Presence Detect State	Read / Write	2
Press Virtual Attention Button	Read Zero, Write One to Set	2
Signal Virtual Power Fault	Read Zero, Write One to Set	2

These fields are more precisely defined in Section 4.3.6.5.

In addition to the above fields, changes by VH software to the Attention Indicator, Power Indicator and Power Controller Control (Purpose 3 above) can generate an interrupt to MR-PCIM.

## 6.1.6. Physical Slot Registers

Each physical slot has an associated set of registers for controlling the Physical Hot Plug Controller. As in PCIe, the Physical Hot-Plug Controller is optional.

If present, the Physical Hot Plug Controller is managed via the Slot registers of the associated Port Table entry.

In addition, when a Bridge Controls Physical Link field in the VS Bridge Table is Set, the Physical Hot-Plug Controller for the Port mapped to that VS Bridge Table entry is also controlled using the Slot Capability, Control, and Status registers of the associated with the Type 1 Header. This is useful

when an MRA Switch is being used as a Base PCIe Switch. It is also useful if MR-PCIM chooses to delegate authority for managing the Physical Link to software running in a specific VH (typically because the associated Port is attached to a Base PCIe Device).

### 6.1.7. Physical Hot-Plug Signals Interface

Hot-Plug support is optional in PCI Express. If provided, some means for communicating Hot-Plug signals from the Switch is needed. In PCI Express, this mechanism is Vendor Specific.

Physical Hot-Plug remains optional in MRA Switches. The Physical Signals Interface also remains Vendor Specific.

The presence or absence of Hot-Plug support is indicated using the Slot Implemented bit in the Port Table. If present, various Hot-Plug signals are optional and their presence is indicated in the Slot Capabilities Register also in the Port Table.

If the Bridge Controls Physical Link field in some VS Bridge Table entry is Set, the Physical Hot-Plug information for the Port mapped to that VS Bridge is also reflected in the associated Type 1 Header.

## 6.2. Virtual Device Migration

Virtual Hot Plug can be used to support Device Migration from one VH to another VH.

To accomplish this, the losing VH gets a Hot Remove sequence. This sequence starts with a push of the Virtual Attention Button and ends with the disabling of the Virtual Power Controller.

Software could then remap and reset the Virtual Device by:

- Ensuring that the gaining VS Bridge is ready to receive the Virtual Device (i.e. the Port / Port VHN fields are unmapped and the Data Link State is clear)

- Clearing the Port / Port VHN fields in the losing VS Bridge Table entry

- Setting the Force Reset bit in the gaining VS Bridge Table entry

- Mapping the Port / Port VHN fields into the gaining VS Bridge Table Entry.

Software would then send the gaining VH a Hot Add sequence. This starts with the assertion of Presence Detect, and finished with the Device being enumerated and used by software in the VH.

## 6.3. Base PCI Express Device Migration

Base PCIe components can also be attached to a Switch and assigned to a single VH. This assignment can change over time through a Device Migration process.

The sequence of operations is similar to that described in Section 6.2. The exception is that since the link is operating in PCIe mode, the Secondary Bus Reset field in the Port Table should be used instead of the Force Reset field from the VS Bridge Table to cleanse Device state for the gaining VH.



## 7. Power Management

MR systems continue to need power management capabilities. Two varieties of power management are involved:

- ❑ Virtual power management allows software running in a VH to believe that it has turned off power to one or more virtual functions.
- ❑ Actual power management allows MR-PCIM software to control device power.

### 7.1. Overview

ASPM and PCI-PM are expanded for MR-IOV. MR Components have both virtual and physical D-states. Slots have virtual and physical power states. Virtual and physical ASPM controls also exist.

### 7.2. Virtual D-State

Every Function in every VH of every MR Device or Switch has a virtual D-state. This includes BFs, PFs, VF's, and Functions as well as P2P Bridges.

Virtual D-state is controlled by software operating in each VH using the rules defined in the *PCI Bus Power Management Interface Specification, Revision 1.2*, and in the *PCI Express Specification*.

Component D-state is an extension of the PCIe multi-function component rules. A MR Component is treated as a multi-function component taking the Functions in all VHs into account. For example, a shared component may not enter L1 state until all Functions *in all VHs* of the component have been written to non-D0 states.

When software in a VH writes a Function to a lower power virtual D-state, the component acts as if it were in a lower power state. Affects on actual power consumption are vendor specific. Vendors are encouraged to use this mechanism to reduce power consumption whenever possible.

Because the component may have not powered down, lower virtual D-states may not result in actual power savings.

### 7.3. Link Power States

Link power states and transitions are unchanged from PCIe. L1 and L2/L3 handshakes are unchanged. As in PCIe, link state is affected by the D-state of all Functions in the Component. In MR, this includes all Functions in all VHs. There are exceptions to deal with D-state of Functions used for managing the MR topology and D-states of Functions in VHs that are not being used.

## 7.4. Multi-Root ASPM

ASPM is a physical notion and is controlled by MR-PCIM using the Port Table (see Section 4.3.3.12 for details). ASPM may also be controlled by the Type 1 Header in each VH if the Bridge Controls Physical Link bit is Set in the associated VS Bridge Table entry.

When the Bridge Controls Physical Link bit is clear, virtual ASPM controls are provided for software compatibility but perform no function. Like PCIe, virtual ASPM controls must support L0s. L1 ASPM support is optional in PCIe and, for simplicity, virtual L1 ASPM is not supported.

## 7.5. Slot Clock and Common Clock Configuration

The Slot Clock Configuration and Common Clock Configuration bits in Type 1 Headers associated with Base PCIe Ports reflect the physical Port.

The Slot Clock Configuration and Common Clock Configuration bits in Type 1 Headers associated with MR Ports reflect the virtual environment and always indicate a common clock. The underlying physical configuration is available in the Port table (see Section 4.3.3.12 for details).

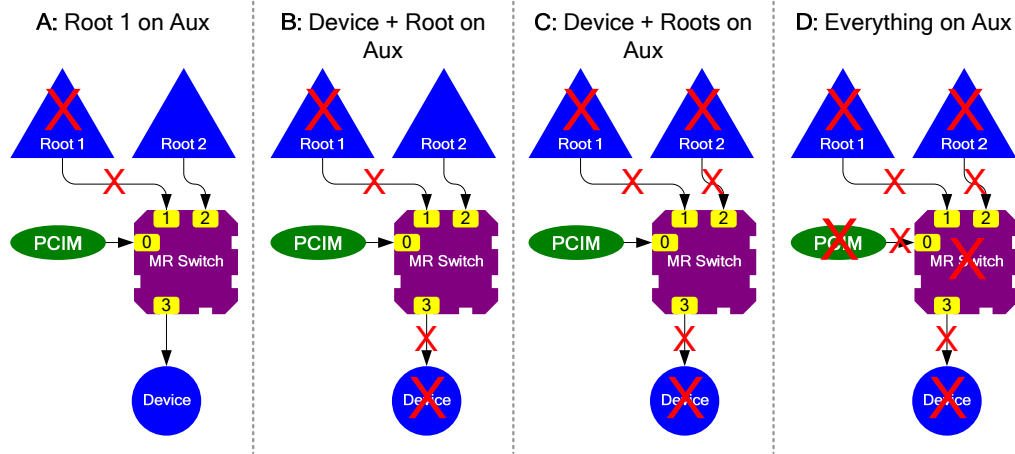
## 7.6. Multi-Root Wake-Up

The wake-up model is mostly unchanged from PCIe. The exception is the need to deal with the “power inversion” situations. In PCIe, power is turned off starting at the leaves of a PCIe hierarchy so that power is removed from a root only after power was first removed from all components below that root. In MR topologies, shared components have *virtual* power removed while the non-shared components have *actual* power removed. This creates situations where a powered on shared component located below a non-shared component (in some VH), needs to wake-up that non-shared component. It also creates situations where a powered off component needs to wake-up a powered on shared component.

Examples of these scenarios are shown in [Figure 7-1](#) below.

Deleted: Figure 7-1





**Figure 7-1: Multi-Root Wake-Up Scenarios**

To address this, MR Switches implement a number of power management features:

- ❑ The ability to send Beacon / WAKE# on receipt of certain PM\_PME messages.
- ❑ The ability to detect beacon / Wake# and convert it into a PM\_PME message.
- ❑ The ability to detect Beacon / WAKE# and signal an interrupt.

#### 7.6.1. PME Triggers Beacon / Wake#

The PME Triggers Beacon / WAKE# bit in the Port Table instructs an MR Switch to issue Beacon or WAKE# when a PM\_PME Message is headed out a Port that is DL\_Down.

**Scenario A:** This mechanism can be used to wake-up Root 1 in [Figure 7-1](#), Scenario A. The MR Device detects a wake-up event in one of its Functions. The link from the Device to the MR Switch is in DL\_Active and the Device is being used by Root 2. The Device sends a PM\_PME Message upstream in the VH associated with Root 1. The MR Switch Port 3 receives the PM\_PME and because it is associated with Root 1, forwards it to Port 1.

Port 1 is DL\_Down and has its PME Triggers Beacon / WAKE# bit Set. This causes the PM\_PME Message to be queued and a wake-up event to be sent to Root 1 (as in PCIe, whether this involves Beacon or WAKE# is platform specific).

#### 7.6.2. Beacon / Wake# Triggers MSI

The Beacon / WAKE# Triggers MSI bit in the Port Table instructs an MR Switch to generate an MSI interrupt to MR-PCIM when it detects a Beacon or WAKE# indication from the Port.

**Scenario B:** This mechanism is used in to power up the Device in [Figure 7-1](#), Scenario B. The Device detects a wake-up event in one of its Functions. The Device is operating on Aux power and generates a Beacon or WAKE# to the MR Switch. The MR Switch interrupts MR-PCIM which

Deleted: Figure 7-1

Deleted: Figure 7-1

notifies the Beacon / WAKE# event and powers up the Device using the Physical Power Controller associated with Port 3. Once the link comes up, Scenario A applies.

**Scenario C:** This mechanism is also used in to power up the Device in [Figure 7-1](#), Scenario C. The Device is powered up as in Scenario B. When the PM\_PME Message is sent, it is addressed to either Root 1 or Root 2 and Scenario A applies to the addressed Root.

Deleted: Figure 7-1

### 7.6.3. Beacon / WAKE# Triggers Beacon / WAKE#

An MR Switch that is operating on Aux power will broadcast a Beacon or WAKE# indication received on a Downstream Port to all Authorized Upstream Ports.

**Scenario D:** This mechanism is used to power up everything in [Figure 7-1](#), Scenario D. The Port containing MR-PCIM is Authorized. The Device is operating on Aux power and generates a Beacon or WAKE# to the MR Switch. The MR Switch generates Beacon or WAKE# out the Authorized Port headed to MR-PCIM. The Root where MR-PCIM is powered on which, in turn, powers on the MR Switch. After the MR Switch is powered on and configured, Scenario C applies.

Deleted: Figure 7-1

## 7.7. Multi-Root PME Turn Off

MR Devices respond to PME\_Turn\_Off messages with PME\_TO\_Ack within each VH. In VH0, PME\_Turn\_Off messages will cause a Downstream Component to request a Link transition to L2/L3 Ready using the PM\_Enter\_L23 DLLP. In non-zero VHs, PME\_Turn\_Off has no effect on link state. Devices must cleanse state in all Functions of a VH before sending PME\_TO\_Ack. This ensures that virtual Device state disappears when the device is “powered off”.

MR Switches process PME\_Turn\_Off messages using PCIe rules within each VS. When a PME\_Turn\_Off message is received at the Upstream Bridge of a VS, the message is broadcast to all Downstream Bridges of the VS. When the downstream sees the downstream component responding with PME\_TO\_Ack, the responses are recorded in a scoreboard. When the last Downstream Bridge responds, a PME\_TO\_Ack message is sent Upstream.

For Upstream Links operating in Base PCIe mode, MR Switches will request the Link to transition to L2/L3 Ready using the PM\_Enter\_L23 DLLP following completion of the PME\_Turn\_Off / PME\_TO\_Ack handshake.

For Upstream Links operating in MR mode, MR Switches will never automatically transition the link to the L2/L3 Ready state. Instead the entry into L2/L3 Ready state is controlled using the Send PM\_Enter\_L23 DLLP bit in the Port Table. See Section 4.3.3.2 for details.

The Port PME Interrupt can be used to determine when all VHs on a Port have completed the PME Turn Off handshake. Bridges that are not mapped or are associated with an Authorized VS do not participate in this determination. See Section 4.3.3.3 for details.

## 7.8. Multi-Root Power Controller

MR-PCIM may enable a virtual power controller for each virtual slot by setting the Virtual Power Controller Present bit in the Hot-Plug Virtual Signals Interface (see Section 4.3.6.5). If this bit is Set

and the Bridge Controls Physical bit is Clear, when the VH turns off power using the PCIe Hot-Plug controller, the VH is sent a reset.

The virtual power controller has no effect on physical power. Turning off virtual power causes the MR Switch to send Reset DLLPs to cleanse state from the affected Components.

A form factor can allow MR-PCIM to control the physical power to a slot. Like in PCIe doing so is optional. This control occurs through the Port Table (see Section 4.3.3.12 for details).

## 7.9. Multi-Root Power Budgeting

Power Budgeting is optional in the PCI Express Specification. Certain form factors may require it. Power Budgeting is required for Devices supporting Hot-Plug. See Section 7.15 of the *PCI Express Specification*.

Power Budgeting remains optional in MR under the same conditions. If provided, power values reflect the power consumed by the BF and all associated PFs and VFs in all VHs.



## 8. Congestion Management

Exceeding the bandwidth of a link or capacity of a buffer can lead to congestion in a topology. In a Multi-Root Topology, congestion may affect the performance of unrelated VHS and lead to Completion Timeout errors. This chapter defines mechanisms for detecting and controlling congestion in a Multi-Root Topology.

### 8.1. Overview

There are three possible causes of congestion in a PCIe topology.

- ❑ A fault in hardware or software configuration of a device in the topology.
- ❑ A static rate mismatch in the capacity of the path from a component injecting traffic into the topology (e.g., a Device) and the ultimate destination (e.g., a Root Port). Congestion due to a static rate mismatch would occur even if the topology were otherwise idle.
- ❑ Traffic merging of multiple flows, none of which individually suffer from a static rate mismatch, causing the capacity of an element in the topology to be exceeded

While the causes of congestion are the same in both single and multi-root PCIe topologies, it is desirable to provide the ability to manage, limit and contain the congestion caused by one VH on other VHS in the system. The congestion management mechanisms outlined in this section allow management of congestion that is unique to MR topologies (i.e., due to traffic merging from different VHS). These mechanisms do not address congestion within a VH since this congestion would have been present in an equivalent PCIe Base topology.

MR congestion management mechanisms provide the following benefits.

- ❑ Preserve the behavior of Virtual Channels (VCs) defined by the PCIe Base specification within a VH.
- ❑ Allow systems to be constructed where a fault in one VH does not result in errors (e.g., Completion Timeouts) in another VH.
- ❑ Allow systems to be constructed that support forward progress guarantees on a VH or groups of VHS when congestion exists on an unrelated VH.
- ❑ Support a wide range of implementation options. At one extreme, they allow creation of MR Devices through incremental changes to SR Devices. At the other extreme, they allow implementations that support complete isolation between virtual hierarchies.

## 8.2. Congestion Isolation

The Virtual Link (VL) mechanism together with Bypass Queues, a logical queuing structure associated with a VL at a Receiver, provide the foundation for isolating congestion and supporting differentiated services within a Multi-Root PCI Express Topology. This section defines these mechanisms and describes them from a system perspective.

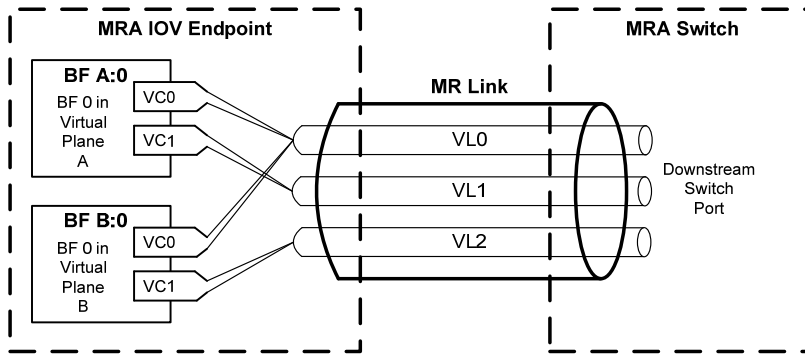
### 8.2.1. Virtual Links

The Virtual Link (VL) mechanism provides the means to support multiple independent logical data flows over a single physical Multi-Root PCIe Link. VLs play the same role in an MR topology as Virtual Channels (VCs) in a PCIe Base topology. VLs are associated with independent fabric resources (queues/buffers and associated control logic) that are used to move information across an MR Link with independent flow control. As with VC in a PCIe Base topology, VLs are associated with a Link and are not end-to-end. Links in an MR topology may implement a different number of VLs.

Multi-Root Aware Components support one or more Virtual Hierarchies (VHs). As defined by the PCIe Base specification, each VH associated with a port of an MRA component may support one to eight VCs. The notation (VH<sub>*x*</sub>, VC<sub>*y*</sub>) is used to denote VC<sub>*y*</sub> associated with VH<sub>*x*</sub>. Each VC of a VH associated with an MRA component port represents an independent logical data flow that in an MR topology must be mapped to a physical VL resource in order for data to be transferred across a Link.

(VH, VC)s may be mapped to VLs in a flexible manner (e.g., (VH 0, VC 0), (VH 1, VC 2), and (VH 2, VC 1) all map to VL 0) or in a VC consistent basis (e.g., (VH<sub>*any*</sub>, VC 0) all map to VL 0). While not a requirement, systems will generally map (VH, VC) data flows with similar QoS characteristics onto the same VL.

A graphical example of mapping VCs associated with VHs to VLs is illustrated in xx, where an MRA device connects to an MRA downstream Switch Port. In this example, the MRA device implements two BFs, each with two virtual channels. One BF in this device has been assigned to VH A while the other to VH B. In this example VC 0 from both VHs has been mapped to VL 0 (i.e., (A, 0) and (B, 0) both map to VL 0) while VC 1 from VH A maps to VL 1 and VC 1 from VH B maps to VL 2. The mapping of (VH,VC)s associated with a port of an MRA component onto VLs is described in section xx.



**Figure 8-1: (VH, VC) to VL Mapping**

#### 8.2.1.1. Virtual Link and Virtual Hierarchy Identification

A port of an MRA component may support from one to eight VLs. Virtual Links are uniquely identified using a Virtual Link Identification (VL ID). There is a fixed one-to-one mapping between VL IDs and VL resources (e.g., VC resource 0 always has a fixed ID of zero (i.e., VL0)).

A port of an MRA component may support from one to 256 Virtual Hierarchies. Virtual Hierarchies associated with a port are uniquely identified using a Virtual Hierarchy Number (VHN). VHNs are Link specific and do not represent a global VH identifier.

Each port is independently configured and managed allowing implementations to vary the number of VLs and VHNs supported per Port based usage model-specific requirements.

MR DLLPs used for flow control accounting contain VHN and VL ID information. Unlike PCIe Base TLPs that contain only TC and no VC information in the header, the MR TLP prefix tag contains both VHN and VL ID information simplifying the Flow Control accounting done at each Port of a Link.

Rules for allocating VL IDs to VL hardware resources associate with a port are as follows:

- ❑ VL ID assignment must be one-to-one
- ❑ The same VL ID cannot be assigned to different VL hardware resources within the same Port.
- ❑ VL ID 0 (VL0) is assigned and fixed to the default VL.

Rules for assigning VH VCs to VL hardware resources associate with a port are as follows:

- ❑ (VH, VC) assignment must be the same (matching) for the two Ports on both sides of a Link.
- ❑ (VH0, VC0) is assigned at initialization, but not fixed, to the default VL.

MR-PCIM is responsible for configuring ports on both sides of an MR link in a consistent manner.

### 8.2.1.2. VL and VC Configuration

Support for VLs beyond the default VL0 is optional. VL0 is always enabled and while not fixed or “hardwired,” by default there is a one-to-one mapping between VC and VLs. Therefore, MR topology initialization may proceed using (VH 0, VC 0) mapped to VL 0 and does not require any specific hardware or software configuration.

MR-PCIM is responsible for enabling VLs and configuring the mapping of VC associated with VHs to VLs.

- ❑ VL0 is always enabled
- ❑ For VLs 1-7, a VL is considered enabled when the corresponding VL Enable bit in the MR-IOV Control register has been set to 1b in the BF, and once FC negotiation for that VL has exited the MRFC\_INIT2\_VL state.

For VLs 1-7, MR PCIM must use the VL Negotiation Pending bit in the MR-IOV Status register to determine when a VL is enabled.

Every VC resource of a VH associated with a Port visible to software operating in the VH must be mapped to an enabled VL. Since the number of VLs supported by components on a Link is implementation specific, and only one VC of any VH may be mapped to given a VL, the number of advertised VC resources to software operating in the VH must not exceed the number of enabled VLs associated with the port.

- ❑ If a function only implements the default VC0 resource, then no configuration is necessary.
- ❑ For Devices this is managed through the Base Function as follows:
  - For Virtual Hierarchies that do not have a MFVC Capability structure associated with the Port, then the VC Extended VC Count field in the Function Control 2 register must be initialized to a value such that the number of VC resources advertised to software operating in the VH is less than or equal to the number of enabled VLs. As a result of this configuration, the VC Low Priority Extended VC Count field in the Function Control 2 register may need to be initialized to a value consistent with the VC Extended VC Count field.
  - For virtual hierarchies that have a MFVC Capability structure associated with a Port, the MFVC Extended VC Count field in the Function Control 1 register must be initialized to a value such that the number of VC resources advertised to software operating in the VH is less than or equal to the number of enabled VLs. As a result of this initialization, the MFVC Low Priority Extended VC Count field in the Function Control 1 register may need to be initialized to a value consistent with the MFVC Extended VC Count field.
- ❑ For Switches this is managed through the corresponding Virtual Switch (VS) Bridge Table Entry as follows:
  - The VC Extended VC Count field in the Switch VS Bridge Control 2 register must be initialized to a value such that the number of VC resources advertised to software operating in the VH is less than or equal to the number of enabled VLs. . As a result of this configuration, the VC Low Priority Extended VC Count in the Function Control 2

register may need to be initialized to a value consistent with the VC Extended VC Count field.

TLPs generated by a VH but not specifically associated with a BF, PF or VF (e.g., interrupts enabled in the Device MR-IOV register) require a mapping to an associated VH. Mapping for these TLPs is performed by the Default VL field in the Device MR-IOV Control register.

TLPs generated by a BF require a mapping to an associated VH. Mapping for these TLPs is performed by the BF VL field in the Device MR-IOV Control register.

### *8.2.1.3. VC to VL Mapping*

A Virtual Link is established when one or more VC IDs from different VHs are associated with a physical resources designated by a VL ID.

Components with Ports that implement VLs beyond the default VL must also implement an associated VC to VL mapping capability. The VC to VL mapping capability is optional for Ports that implement only the default VL0. Ports that do not have an associated VC to VL mapping capability must map VC0 from all supported VHs to VL0.

In order to preserve the semantics of a VC defined in the PCIe Base specification, only one VC of a given VH may be mapped to a VL. The behavior when two or more VCs from the same VH are mapped to a single VL is undefined.

Given the above requirement, knowledge that a VH is mapped to a VL together with the VC to VL mapping function associated with that VH is sufficient to determine the (VH,VC). Thus, indicating a that a VL has a mapped VH, or (VL,VH) is synonymous with specifying the (VH, VC).

VC to VL mapping is from virtual VC IDs to VLs and is controlled as follows:

- ❑ For Devices, the VC to VL mapping is controlled by fields in the Function VC to VL Map register associated with the BF of each VH. This map is from Virtual Hierarchy VC resources as they would have appeared on the Link in an equivalent PCIe Base component. Thus, if function 0 in the VH contains a MFVC Capability structure then this mapping is from VC IDs managed by the MFVC Capability structure. Otherwise, this mapping is from VC IDs managed by the BF VC capability structure.
- ❑ For Switches the VC to VL mapping is controlled by fields in the VC to VL Map register in a VS Bridge Table entry.
- ❑ A VC ID  $x$  is mapped to a VL  $y$  when the VC $x$  VL Map field has been initialized with a value of  $y$  and the corresponding VC $x$  VL Map Enable bit has been set.

VC to VL mapping may be performed during MR-PCIM initialization or on an as-needed basis as dictated by software operating in the VH. The PCIe Base specification supports the arbitrary mapping of VC IDs to VC resources. Thus, in general, MR PCIM has no a priori knowledge of which VC IDs will be used or how they will be allocated by software operating in the VH. In systems where MR PCIM possess this knowledge or in which MR PCIM can communicate desired allocation to software operating in the VH, VC IDs may be mapped to VLs during MR-PCIM initialization (i.e., prior to the instantiation of software operating in the VH). Otherwise, this allocation must be performed by MR-PCIM on an as-needed basis as VC IDs are allocated by software operating in the VH to VC resources.



The VC ID to VC resource allocation and TC to VC map configuration performed by software operating in the VH may be determined by MR-PCIM via polling or interrupts as follows:

- ❑ For Devices this information may be determined by examining the Function Table Resource Status and Function Table Multi-Function Resource Status registers. Any modification of these fields can be used to trigger a BF interrupt to MR PCIM.
- ❑ For Switches this information may be determined by examining the VC Resource Fields register. Any modification of these fields can be used to trigger a VS Bridge interrupt to MR-PCIM.

If software operating in the VH enables a VC that has not been mapped to a VL, then the VC Negotiation Pending bit in the VC Resource Status Register remains set until the VC ID has been mapped by MR-PCIM to a VL. Once mapped to a VL, the VC Negotiation Pending bit is cleared.

Once mapped, VC to VL mapping may only be modified when a VL and all associated VCs are disabled. Prior to disabling a VL, VCs mapped to that VL from all virtual hierarchies must be disabled. The behavior of disabling a VL with mapped and enabled VCs is unspecified.

Mapping a VC to a disabled VL results in the VC Negotiation Pending bit in the VC Resource Status Register remaining set (i.e., the VC resource does not complete the process of negotiation due to an invalid mapping). If at some later point the disabled VL becomes enabled, the state of the VC Negotiation Pending bit is undefined.

Once enabled, MR flow control information is tracked by Receivers and Transmitters for all configured VLs and (VH, VC)s. The enabling or disabling of a VC within a VH represents a logical event that does not affect the operation of a physical MR Link or flow control information tracked on that Link. For example, (VH,VC) flow control continues to be tracked during a hot-reset of a VH.

#### *8.2.1.4. Arbitration*

The objectives of MR arbitration are to provide the following:

- ❑ Guaranteed forward progress on all supported data flows.
- ❑ Differentiated service characteristics for data flows associated with a VL within an MR topology.
- ❑ The ability to tune bandwidth and end-to-end latency between components in an MR topology.

MRA components that support multiple VHs require an arbitration mechanism associated with each supported VL at egress Ports to select the VH from which the next TLP will be transmitted on that VL. This arbitration is referred to as VH to VL arbitration.

MRA components that support multiple VLs require an arbitration mechanism at egress Ports to select the VL from which the next TLP will be transmitted on the physical Link. This arbitration is referred to as VL to Link arbitration or just VL arbitration.

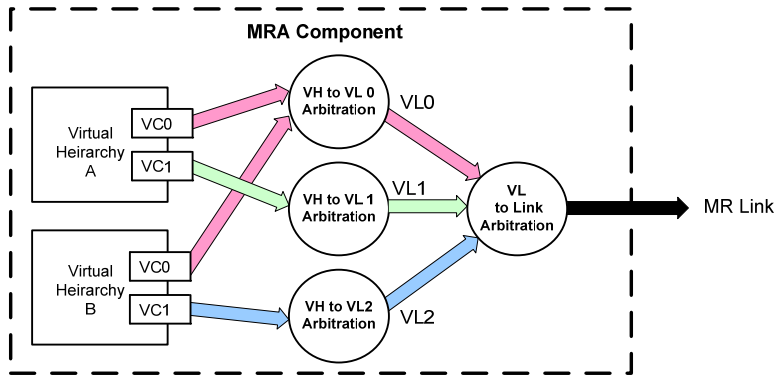


Figure 8-2: MRA Arbitration Model

Figure 8-2 illustrates both VH to VL and VL to Link arbitration associated with an egress Port of an MRA Enabled PCIe Component such as an MR Root Port, Switch, Bridge, or Device. The component in this example implements two VHs, with both VHs implementing two VCs. Flows (A, VC 0) and (B, VC 0) are mapped to VL 0 and require VH to VL arbitration to control the multiplexing of TLPs onto this VL. VLs one and two only have a single mapped flow and therefore require trivial arbitration. The physical link associated with the egress Port in this example implements three VLs. VL to Link arbitration is required to control the multiplexing of VLs onto the physical link.

#### 8.2.1.4.1. VH to VL Arbitration

MRA components that implement multiple Virtual Hierarchies must implement a VH to VL arbitration mechanism associated with each VL supported by a Port.

VH to VL arbitration is not configurable. All implementations must support a hardwired-fixed VH to VL arbitration scheme (e.g., Round-Robin) that guarantees forward progress on all VHs associated with a VL at an egress Port.

#### 8.2.1.4.2. VL to Link Arbitration

MRA components with ports that support multiple Virtual Links must implement a VL to Link arbitration mechanism.

A component may support a hardwired-fixed arbitration algorithm or optional software configurable algorithms selection. Support for the optional software configurable arbitration algorithm selection is indicated by the state of the VL Arbitration Table Present bit in the MR-IOV Capabilities register.

If an implementation does not support software configurable algorithm selection, then it must implement a hardwired-fixed arbitration scheme (e.g., Round Robin) that guarantees forward progress on all enabled VLs.

The remainder of this section describes the behavior and requirements of software configurable VL arbitration algorithm selection.

VL arbitration algorithm selection is controlled as follows:

- ❑ For Devices, registers associated with VL arbitration are located in the MR-IOV Capability.
- ❑ For Switches, registers associated with VL arbitration are located in the Port Table Entry of the corresponding port.
- ❑ The VL arbitration table in all components, if present, is located in BAR memory space.

VLs may be partitioned into two priority groups – a lower and an upper group. VLs in the upper group are arbitrated using strict priority based on VL number while VLs in the lower group are arbitrated only when there are no packets to process in the upper group. Arbitration within the lower group may be configured to one of the supported arbitration algorithms described below.

Membership of a VL in the low or high priority group is determined by the state of the corresponding bit in the VL Strict Priority Arbitration field in the VL Arbitration Control register. Since the VL Strict Priority Arbitration field represents a bit vector, VLs to group assignment is flexible and need not be allocated sequentially based on VL ID.

Among VLs configured for strict priority, priority is based on increasing VL number. VL0 has the lowest priority while VL 7 has the highest.

The arbitration algorithm for VLs in the low priority group is selected by the VL Arbitration Select field in the VL Arbitration Control register. Arbitration algorithms supported by an implementation are advertised in the VL Arbitration Capability field in the VL Arbitration Capability and Status register and may include the following architected schemes.

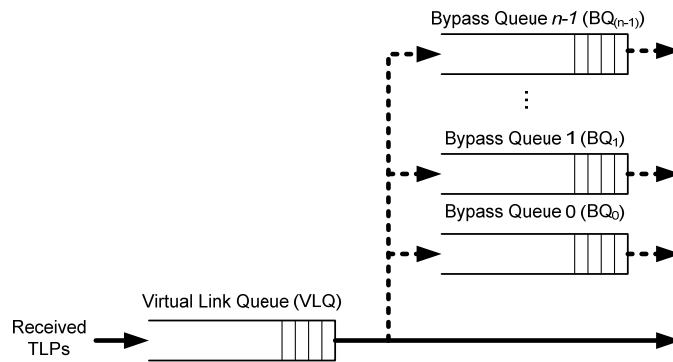
- ❑ Hardware-fixed arbitration, e.g. Round-Robin
- ❑ Weighted Round Robin (WRR) arbitration scheme with 32, 64, 128 or 256 phases
- ❑ Time-Based Weighted Round Robin (time-based WRR) arbitration scheme with 128 phases
- ❑ Vendor defined arbitration

This specification establishes a standard framework within which vendors may specify their own vendor specific arbitration scheme. The definition of vendor-defined arbitration is outside the scope of this document.

VL arbitration algorithms, e.g., WRR and time-based WRR, operate in a manner analogous to the schemes defined for VC arbitration in the PCI Express Base specification.

### 8.2.2. Bypass Queues

Associated with each VL at a receiver are dedicated physical resources (queues/buffers and control logic) that allow independent traffic flows to proceed on the Port. This section describes the logical queuing structure associated with each VL at a Receiver. This logical description does not imply or require a particular implementation and is used solely to clarify requirements.



**Figure 8-3: Logical Queuing Structure Associated with a VL Receiver**

As illustrated in [Figure 8-3](#), associated with each VL at a receiver is a Virtual Link Queue (VLQ) and one or more optional Bypass Queues (BQ). Within a VL, the PCIe ordering rules defined in the Base specification are maintained with respect to TLPs from a VH across both the VLQ and the associated BQ (if any). TLPs associated with a VL received on a Link are queued in the VLQ. In the absence of congestion, TLPs for non-congested VHs are dequeued from the VLQ for receiver processing.

Deleted: Figure 8-3

When a TLP at the head of the VLQ experiences congestion, a free BQ is allocated to the VH associated with the TLP, if one does not already exist, and the TLP is moved to the BQ. Subsequent TLPs associated with the VH from the VLQ are moved to the same BQ until the congestion condition clears and the (now empty) BQ is freed. This scheme allows TLPs associated with a congested flow to be bypassed and isolates congestion to a VH. TLPs are dequeued from a BQ as flow control credits of the required type for the VH associated with the BQ become available. An implementation must ensure that forward progress is guaranteed on all flows.

If at any point the number of congested VHs associated with a VL exceeds the number of BQs implemented by a Receiver for that VL, then TLPs in the VLQ experience congestion. Thus, congestion is managed within a VL at the receiver by dynamically mapping congested VH flows onto BQs. The degree of congestion isolation between flows mapped to the same VL is dictated by the number of BQs implemented for that VL by the receiver. A receiver that implements only the required VLQ provides no congestion isolation between data flows mapped to the VL, while a receiver that implements  $n$  optional BQs provides complete congestion isolation for up to  $n$  flows.

### 8.2.3. Flow Control Rules

- ❑ Components must implement independent Flow Control of all supported VLs.
- ❑ As in the PCI Express Base specification, flow control distinguishes between TLP type (Posted, Non-Posted, and Completion) and Header/Data. Thus, there are six types of tracked flow control information.
- ❑ The unit of Flow Control credit is 4 DW for data

- ❑ The unit of Flow Control credit for headers is one maximum-size header plus TLP prefix and TLP digest
- ❑ Flow Control is initialized autonomously by hardware only for the default virtual link (VL0) and virtual hierarchy (VH0).
- ❑ When other Virtual Links are enabled by MR-PCIM, each newly created VL will follow the flow control initialization protocol.
- ❑ (VL, VH) credits are negotiated using the flow control initialization protocol outlined in Section 2.1.2 whenever MR-PCIM increases NumVH.
- ❑ A Receiver must never cumulatively issue more than 2047 outstanding unused credits to the Transmitter for data and 127 for header.
- ❑ If an Infinite VL and (VL, VH) credit advertisement has been made during initialization, no Flow Control updates are required for that VL following initialization.
- ❑ A Receiver that advertises non-infinite VH credits must utilize MRUpdateFC DLLPs for that VL.
  - Independent MRUpdateFCs DLLPs are used to track header and data credits associated with VLs and (VH,VC)s.
  - As described in Section tbd, Receivers and Transmitters track independent flow control information for each VL and for each supported VH. For each VL and (VL, VH), the six types of flow control information outlined above are tracked.
  - A TLP in a Receiver's VLQ or BQ consumes both VL and corresponding VH credits.
  - Both VL and corresponding VH credits are released when a TLP is processed and removed from the logical queuing structure associated with the Receiver for a VL.
  - MRUpdateFC DLLPs are only associated with VH credits related with a VL. VL credits are implicitly computed using state information and MRUpdateFC DLLPs as outlined in Section 2.4.1.
- ❑ A Receiver that advertises non-infinite VL credits and infinite VH credits must utilize PCIe Base UpdateFC DLLPs for that VL.
  - UpdateFC DLLPs are used to explicitly track VL header and data credits.
  - The VH gating function is unconditionally satisfied for all Credit Types associated with that VL.
  - Receivers and Transmitters track independent flow control information for each VL. For each VL, the six types of flow control information are tracked.
  - A Receiver that advertises infinite VH credits may only implement a VLQ for that VL.
  - VL credits are released when a TLP is processed and removed from the logical queuing structure associated with the Receiver for a VL.

- ❑ If a receiver advertises infinite VH credits on one VL, then it must advertise infinite VH credits on all VLs.

### 8.3. Performance Monitoring and Statistics Collection

Congestion in a Multi-Root Topology may result in traffic associated with one Virtual Hierarchy affecting the performance of an unrelated Virtual Hierarchy. This section defines a set of optional performance monitoring and statistics collection capabilities that may be used to diagnose, plan and tune the performance of a multi-root system.

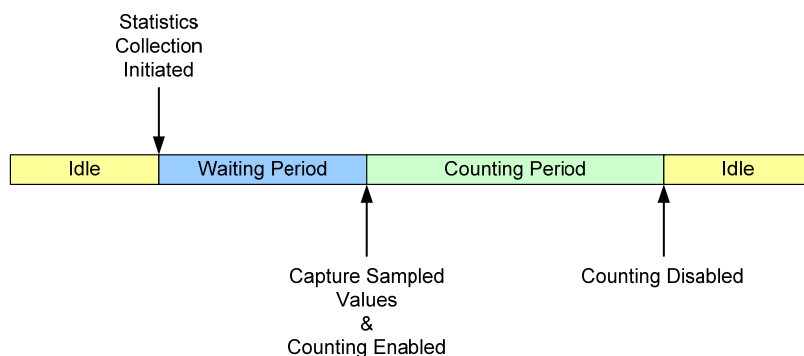
The objective of standardizing these capabilities is to allow component vendor independent software to monitor performance and manage congestion. It is not a goal of these capabilities to monitor or count errors.

This capability is optional for all MR-IOV components, but its implementation is strongly encouraged for all MRA Switch Ports.

Registers and tables associated with the Performance Monitoring and Statistics CollectionsCollection capability are described in Section 4.5. This section describes functional behavior.

A component that implements the optional Performance Monitoring and Statistics Collection Capability is required to implement basic features outlined in this section and in Section 4.5. These features ensure minimum functionality and interoperability with software that utilizes these capabilities.

Support for the Performance Monitoring and Statistics Collection Capability is indicated by a non-zero value in the Statistics Capability register location in the MR-IOV Extended Capability.



**Figure 8-4: Statistics Collection Process**

Performance statistics are recorded by Statistics Counters and may be captured values and counted values. Captured values correspond to sampled system state while counted values correspond to the number of occurrences of a selected event over a counting period. The statistics collection process is illustrated in [Figure 8-4](#).

Deleted: Figure 8-4

Completion of the statistics collection process (i.e., the end of the counting period) may be signaled via an interrupt.

A component that implements the Performance Monitoring and Statistics Collection Capability must implement a Statistics Block Table and a Statistics Descriptor Table. These tables are located in memory space and their location is specified by the Statistics Descriptor Table and Statistics Block Table registers in the MR-IOV Extended Capability structure.

A set of Statistics Counters that share a common initiation mechanism and statistics collection process periods is referred to as a Statistics Block. A component that implements the Performance Monitoring and Statistics Collection Capability may implement one to 32 Statistics Blocks. Each Statistics Block has an associated Statistics Block Table entry that contains a pointer to a Statistics Counter Table that holds the Statistics Counters associated with the Statistics Block. The Statistics Block Table entry also specifies the statistics collection process state (i.e., Idle, Waiting, Counting), number of entries in the Statistics Counter Table, waiting period, and counting period.

Statistics Counters associated with a Statistics Block may have different characteristics and be associated with different Ports. Associated with each Statistics Counter is a Statistics Descriptor Index that points to the Statistics Descriptor Table entry that describes statistics that may be recorded by the Statistics Counter. The actual statistic recorded by a Statistics Counter is selected by the Statistics Select field.

Associated with a Statistics Counters is a 64-bit counter that is used to report the captured statistic. The counter is formed by the Statistics Counter Low and Statistics Counter High registers. While the field is specified as 64-bits, an implementation is free to implement fewer bits. The number of implemented counter bits is specified by the Statistics Width field and, for standard counters, must be 32-bits or greater.

Associated with each Statistics Counter is an optional filter specified by the Statistics Filter Enable and Control register. Filters allow refinement in a recorded statistic. For example, rather than count all transmitted TLPs on a Port, a filter may be used to only count transmitted TLPs from a particular VH on a particular VL. Each entry in a Statistics Descriptor (i.e., an S-bit) defines required filters that must be implemented and optional filters that may be implemented.

The format of Statistics Descriptors, standard statistics, filters and requirements are specified in Section 4.5.2.

A component that implements the Performance Monitoring and Statistics Collection Capability must implement at least one Statistics Block and at least two Statistics Counters per Port. At least two Statistics Counters per port must implement Statistics Descriptor standard statistics specified as required in [Table 4-69](#).

Deleted: Table 4-69